THE FACULTY OF SOCIAL SCIENCES,

DEPARTMENT OF INFORMATION SCIENCE AND MEDIA STUDIES

Improving Image Retrieval with a Thesaurus for Shapes

The VORTEX Prototype

THESIS

Submitted in partial fulfilment of the requirements for the degree of

Candidatus rerum politicarum

By Lars-Jacob Hove

November 2004

Acknowledgements

There are several people that have contributed to the work encompassing this thesis. Above all, I would like to thank my supervisor, Associate professsor Joan C. Nordbotten. Her guidance and support has proved invaluable both in the start phase and throughout the entire project.

I would also like to thank Professor Emeritus Svein Nordbotten for taking the time to give me good advice during my work with this project.

A big thanks is also in order to my fellow student Amund Trovåg for meaningful discussions and contributions to the work described in this thesis. I would also like to thank Arne Helgesen and Aleksander Krzywinski, also fellow students, for contributing to keep our academic motivations at a reasonable level in our shared office. My gratitude also goes to Tormod Skauge, at the Institute for Chemistry, for his offers of guidance with the statistical methods used in the project. I would also like to thank the fellow students working in the *Virtual Exhibit on Demand* project for valuable assistance, intersting discussions and suggestions during the last two years.

I would also like to offer my thanks for the support my family have given me before and during this project.

Finally, I am also very grateful for the support and understanding I have received from Anne. I would not have managed to finish this project without your support.

Table of contents

List of Figures

List of Tables

1 The Challenge of Semantic Image Retrieval

The focus of this thesis is how to achieve successful *semantic image retrieval from large scale image collections*. We explore some of the difficulties of getting relevant results when searching through image collections. A novel tool, *The Shape Thesaurus,* is proposed as a step towards improving image retrieval.

Images have been used as a means to convey information since the dawn of man, either alone or in conjunction with written text. Today, images are used in almost every walk in life, and a diverse range of professional groups make use of images. The police use visual information to identify people and record crime scenes as evidence. In medical and health professions, images from the visual spectrum as well as x-rays and ultrasound are used for diagnosis and monitoring of patients. Photographs are used in architectural and engineering design to record and document finished projects, and they are very often used in combination with text, as they are capable of conveying information difficult to describe in words. Finally, they can be appreciated in their own right, either as works of art or personal souvenirs.

Until the previous decade, images were prevalently stored as physical objects. Limitations in both hardware and software made computers an ill-suited tool for image collections. However, as the computational power of both hardware and software have increased, the ability to store more complex data types, such as images, in databases, has been drastically improved.

1.1 Information Retrieval, Image Retrieval and Image Processing

The theoretical framework for this thesis comes from several different research areas. It has its theoretical foundations in the field of *Information Retrieval*, an important research area within the field of Information Science. According to Baeza-Yates and Ribeiro-Neto (1999), it concerns *"the representation, storage, organization of, and access to information items"*. While Information Retrieval originally focused on text based information items, it has grown to accommodate "new" digitalized information items, such as video, sound and images. Furthermore, the thesis has most of its theoretical foundations in the field of *Image Retrieval*. This is a relatively young field of research, and may in many ways be regarded as a part of *Information Retrieval*. Research started in the 70s, based in traditional Information Retrieval. However, since then, research in the field has been driven by both the fields of Information Retrieval and *Computer Vision*. Today it is very active and important research area, spanning a broad range of research disciplines, such as Information Retrieval, Cultural Studies, Computer Vision and Image- and Signal Processing.

Many of the tools, techniques and methods used in this thesis, and in Image Retrieval as a whole, come from the field of *Image Processing*. *Image Processing* refers to a computer discipline wherein digital images are the main data object. It covers the analysis, manipulation, storage, and display of graphical images from sources such as photographs, drawings, and video. This type of processing can be broken down into several sub-categories, including: compression, image enhancement, image filtering, image distortion, image display and colouring, image analysis and comparison, and image editing.

A closer look at the current literature and research available in the Image Processing community reveals that a lot of effort is being put towards creating and improving efficient tools, techniques and algorithms for narrow and specialized domains. For application domains with homogeneous image collections, such as face- or fingerprint identification, or areas in medical image processing, very efficient algorithms are now available.

However, although efficient within their own domain, the techniques developed for specialized areas are unlikely to provide support for more general application areas, and more effort is still needed towards developing tools for *Visible Image Retrieval* applications, or VIR. VIR is concerned with retrieval of images from heterogeneous collections of single images generated with visible spectrum technologies (Colombo and Del Bimbo 2002). Images can be taken using equipment capturing different frequencies of electromagnetic radiation, from Gamma- and X- rays, through the visible spectrum to AM and Long Wave radiation. The *Visible Spectrum* is part of the electromagnetic radiation we know as *Visible Light*. In other words, what we ordinarily perceive as "images", excluding images captured with X-ray, infrared light and so on. VIR is the focal point of this thesis.

Although this thesis is based on theory from all the above mentioned disciplines, it is rooted in Information Retrieval, and the assumptions and definitions presented here might differ from what one would find in a thesis based in computer vision or signal processing.

1.2 The Challenges of Current Image Retrieval Systems

Current available image collections and image databases are, to a large extent, based on keyword annotation for image indexing and retrieval. In systems such as these, images are annotated with descriptive keywords, a process which requires a lot of manual work, especially in larger systems. In addition, different people may perceive the contents of an image differently. An art student will most likely describe a baroque painting differently than a high school student. In addition, different people's understanding of the semantics of a certain keyword might wary (Rui, Huang et al. 1998). These are known are the problems of *Volume* and *Subjectivity*. During the previous decade, *Content Based Image Retrieval*, or CBIR, emerged from the field of Computer Vision as a possible solution to these problems.

CBIR systems consist of automatic indexing methods and search and retrieval techniques for description and retrieval of images based on their content. Current CBIR mechanisms can, to a certain degree, successfully compare and retrieve images based on syntactical features, such as colour, texture, shapes and spatial placing of objects within images. However, automatic retrieval of images based on higher level content, such as their semantic content, has proven difficult. While retrieval based on syntactical features has proven to be enough for some domain areas, we want, and need, retrieval systems capable of indexing and retrieving images based on higher level content. This gap between what is possible in the currently available technology, and what we want, has been dubbed *the Semantic Gap*. An illustration of this can be found in Figure 1, below.

Figure 1 - Images depicting dolphins in various poses.

All three images depict a single dolphin. The first two images shows slightly arced dolphins, beak to the left, fin at the top and tail to the right. Although one is a drawing and one is a photographic image, both have very similar shape properties, and we clearly see how they are similar on a structural level. However, while the last image is quite similar to the two first images in semantic content, there are few similarities on a syntactical scale, except maybe in colour. While we could expect that most current CBIR mechanisms would be able to find the similarities between the two first images. However, it is unlikely that it would find any likeness between the two first images and the last, expect maybe in colour.

We want an image retrieval system to be able to retrieve images which are similar both in *structural* and *semantic* content. Much of the current research in image retrieval is aimed at bridging the semantic gap, or making the gap smaller. The framework suggested in this thesis is an effort towards this goal.

1.3 Proposed framework – the Shape Thesaurus

The main motivation behind this thesis has been to evaluate the possibilities of narrowing the Semantic Gap; is it possible to create a framework which can be used by an image retrieval system to assist in retrieving images with similar *semantic* content but differing *structural* content? In this thesis, a novel approach to this question is suggested by borrowing ideas from text based information retrieval, and combining these with image analysis and comparison techniques from image processing. The resulting structure, *the Shape Thesaurus*, is proposed as a possible extension to an image retrieval system. The shape framework is based upon similar use of thesauri in image retrieval applications. The framework is discussed in detail in chapter 3.

Figure 2 below shows a possible architecture for a standard image retrieval system expanded with a shape thesaurus. The white boxes represent the components we could expect to find in most image retrieval systems with support for CBIR. The grey boxes represent the additions of the shape thesaurus. We see that the shape thesaurus acts as a supplementary tool to the existing components of an image retrieval system. It is intended that the framework described here can be added to both existing and new image retrieval systems.

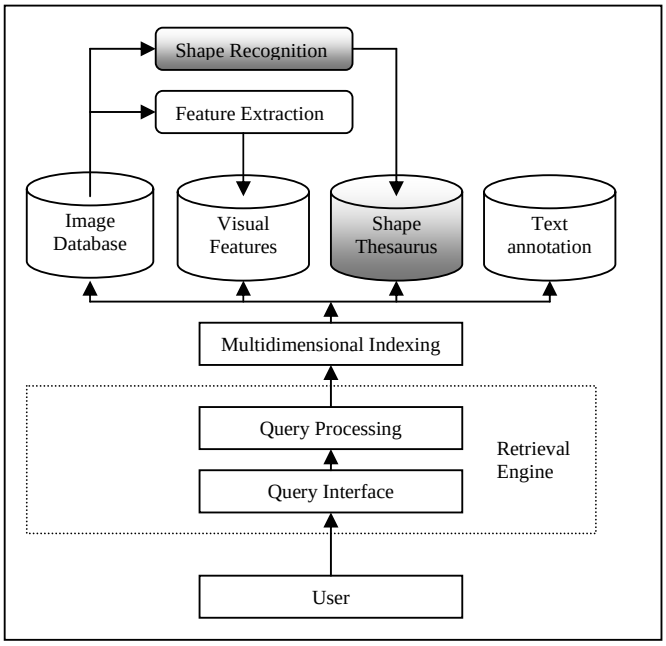


Figure 2 – Proposed architecture for an Image Retrieval system. Based on Huang and Rui (1999).

Figure 2 also illustrates the boundaries of the thesis. Even though an image retrieval system enhanced with a shape thesaurus should consist of all these elements, the goal of this thesis has been to test and evaluate the properties of the shape thesaurus itself. Although it would be very interesting to evaluate the capabilities of a complete, thesaurus-extended image retrieval, this has been outside the scope of this thesis.

1.4 Research Project

The main goal of this research project is to evaluate the possibilities presented by the Shape Thesaurus framework described in this thesis. The project is regarded as a pilot study, and the findings in this thesis can be used as a foundation for further studies of this, or similar, approaches to image retrieval.

1.4.1 Research Question and Hypothesis

The research project described here has been formulated into the following research question:

> *Can recall / precision measures for an image retrieval system be significantly improved by utilizing a thesaurus for shapes?*

The following hypothesis is proposed in order to investigate the research question:

> *An image retrieval system that utilizes a thesaurus for shapes, will lead to a significant improvement in recall / precision results over a system based on syntactical feature comparison.*

Measurement of *recall and precision* is a method for examining the quality a search (Baeza-Yates and Ribeiro-Neto 1999; Lu 1999). This measurement was used as it is a widely known measurement for information retrieval systems. It is relatively

straightforward to compare different retrieval techniques using this method, making it a suitable tool for evaluating the hypothesis and research question.

1.4.2 Methodological Approach

The project can be classified as *design research*. A prototype shape thesaurus has been built, and evaluated in experiment, in an attempt to gain information about the framework. The prototype has been developed using Oracle 9i interMedia (O9i)[1], and is part of a prototype image retrieval system named VORTEX; Visual Object Retrieval – Thesaurus EXtension. The implementation of this prototype is described in chapter 4.

According to the research question and main hypothesis, an image retrieval algorithm based on a Shape Thesaurus has been compared to an algorithm based on syntactical features for retrieval. As there was no readily syntactical feature based retrieval systems available, a basic CBIR algorithm was implemented as part of the VORTEX system, using the built-in CBIR functionality of O9i. In addition to being relatively straightforward to implement, O9i represents one of the two major database management systems with extended support for images. The term OCBIR is used throughout this thesis to refer to this retrieval algorithm.

Both retrieval algorithms have been measured using recall and precision using a set of 24 queries. The queries sets were divided into 4 levels of complexity, describing different levels of image content: generic objects, generic scenes, specific objects and scenes and finally abstract content. Queries based directly on syntactical features were not tested, as the proposed framework is aimed at retrieving images based on their *semantic* content. The queries were developed based on literature described in chapter 2. An overview of the queries can be found in Table 8, in chapter 5.2.2.

The queries were expressed as both *example images* and *sketches*, representing the two major approaches to visual image retrieval; *Query-By-Example* and *Query-By-Sketch*. The 24 queries were given to a group of 6 persons who supplied the sketches and example images used as input to the two retrieval algorithms. Although it would be possible for me to create the input images myself, this would possibly bias the results as I had full knowledge of the images in the image collection. The test group was thus included in the experiment in order to reduce this bias. The resulting seed images are available in appendix G.

The seed images were used as query input to the two retrieval algorithms, and the results were measured using recall and precision a tool. Three different recall and precision measurements were measured and used as a basis for comparison between the two algorithms; *recall and precision curves*, *recall and precision histograms* and *single value summaries*. Furthermore, the single value summaries were used in a *Student-T* test to examine if any *significant* differences could be found between the two algorithms.

The results from the individual queries can be found in Appendix I, the recall and precision measurements in Appendix J, while the results of the significance testing can be found in Appendix K. A discussion of these results can be found in chapter 6.

[1] The abbreviation O9i is user for Oracle 9i interMedia throughout this document.

In addition to the queries, the respondents were given a questionnaire in order to collect data about how they felt about using images as visual expressions of a query, and how difficult they found it to be. This was not central to this thesis, and only limited effort has been put towards evaluating the questionnaire. There was a small group of respondents (5), and a small set of questions. However, the findings are interesting, and might be used as a seed for further research into the user-end of image retrieval systems. The results of the survey and the questionnaire are briefly discussed in chapter 5.3.5. The results can be found in appendix I.

Finally, the Shape Thesaurus framework described in this thesis was used as a basis for a paper presented at the conference NOBIM 2004 (Norwegian Conference on Image Processing and Pattern Recognition) (Hove 2004). Furthermore, a paper focusing more on the implementation and testing of the shape thesaurus in the VORTEX prototype has been accepted for presentation at NIK 2004 (Norsk Informatikk Konferanse) (In reference).The two papers are available in appendix L.

1.4.3 Experimental Results

The results from the evaluation of the two algorithms indicated that the inclusion of a shape thesaurus might be beneficial to an image retrieval system. The evaluation showed that the recall values for image retrieval with a shape thesaurus were significantly higher than the recall values achieved by OCBIR. While the experimental results have to be taken with certain reservations, they indicate that the principles presented in this thesis is worthy of further investigation.

1.5 Limitations

The field of Image Retrieval is a very large field. It has neither been possible nor the intention to cover all possible faucets of image retrieval. Henceforth, this project focuses solely on the *search results* achieved by the two systems. Search efficiency in terms of execution time and similar measurements has not been evaluated. Neither have such important areas as image compression techniques, different image formats and issues relating to these been taken into consideration in this thesis.

The VORTEX prototype has not been developed into a fully working image retrieval system, and the shape thesaurus implemented as part of the prototype does not contain all the suggested functionality presented in this thesis. Only the basic functionality needed to evaluate the fundamental principles of a shape thesaurus has been implemented.

1.6 Appendix

The appendices referred to in this thesis are available in two different versions. Digital versions are available on the CD accompanying this document. Furthermore, a hardcopy version of the appendix is available as a separate document, also accompanying the main document.

2 Theoretical framework and Literature Review

2.1 Information, Media and Imagery – Fundamental Concepts

What is *information*? How is it represented and interpreted? How is knowledge transferred between individuals, and what role do images have in this? We live in what has been dubbed *The Information Age*, and information has become one of the most important resources to companies and individuals alike. Companies worldwide use vast amount of money and resources on *information management systems*, and the economics of most industrialized countries are to a high degree based on trade and development of information. Words like *Knowledge, Information, Data, Databases, Multimedia,* and so on, are used everywhere. Companies use them as buzzwords for their latest products, they are used interchangeably in literature and the semantics of the words have become blurred. A clear grasp of these basic concepts is fundamental for any discussion in this field, and in the following text, and throughout this thesis, definitions are given for these basic concepts.

2.1.1 Knowledge, Information and Data

We start our discussion with definitions of some of the central concepts in Information Retrieval.

Definition 1[2] **-** ***Knowledge*** *is familiarity, awareness, or understanding gained through experience or study, the sum or range of what has been perceived, discovered, or learned.*

The ability to store knowledge for future use is one of the fundaments for development and progress. Without it, knowledge has to be passed from person to person through oral communication, and some of the accumulated knowledge of an individual is lost when that individual dies. Much knowledge has survived thanks to oral communication, for instance folk music, folk tales and so on. The form was more important than the content, which was difficult to record prior to the 19th century. Today we accumulate much more data and the passing on is no longer a person-to-person action, but a continuous flow of multiple passing on. This is beyond the capacity of the earlier time's method and media. The ability to store data and makes it possible to share knowledge with others as information, and allows later generations to build upon this. This leads us to the next set of definitions:

Definition 2 - ***Data*** *are symbols inscribed in formalized patterns, representing facts, observations and/or ideas that are capable of being communicated, interpreted and manipulated by some human or mechanized process.* (Nordbotten 2004).

Definition 3 - ***Information*** *is the meaning that a human expresses by or extracts from data by means of known conventions of the representation used. (Gould 1971)*

[2] Definitions of important terms are given throughout this document. The definitions are collected in appendix A. Unless otherwise stated, the definitions have been put together by the author, based on different sources and lexicographical definitions.

While this gives a definition of the relationship between knowledge, information and data, it is necessary to point out that data is not exclusively a representation of factual knowledge; it might just as well represent stories, messages and ideas. The definitions above also point out a very important issue; for *data* to become *information*, it is necessary for a human to interpret it. The interpretation is more often than not biased by cultural context; interpretation is subjective and dynamic, and a process that may give different results every time. This has two implications. Firstly, in order to decode the data, it is necessary for the interpreter to have knowledge of the language, or protocol, used for encoding. If the interpreter lacks this knowledge, the data gives no meaning, as was the case with Egyptian hieroglyphs until the discovery of the Rosetta Stone. Furthermore, it implies that the interpreter is capable of a higher understanding of the message encoded in the data.

2.1.2 What is an Image?

With these fundamental concepts in place, it is possible to discuss images and their role in human knowledge and its representation. Ever since the early humans created cave paintings depicting hunting scenes, man has been using images to convey information. The old proverb goes "An image says more than thousand words". This is true because images are rich in information, and can be used by people from a broad range of disciplines. Consider a set of photographs depicting a busy street scene a century ago. Historians may find the scene useful as a snapshot of the times; an architect can find information about buildings and structures, while cultural historians can study changes in fashions.

The word "Image" stems from the Latin word *imago* (imitation, copy, likeness, bust), but an image is generally a representation, or double of something. In common usage, it is an artefact that reproduces the likeness of some subject, at several different levels. At the most basic level, they represent a response to light, while at the most complex level they represent abstract ideas dependent on the viewers knowledge, experience and mood. Note that the term "image" is formally a broad term; including images captured using the whole of the electromagnetic spectrum, such as x-rays and ultrasound. However, this thesis is limited to the study of images captured using *visible light* or *visual spectrum technologies*[3]. Based on this, the following definition is proposed:

Definition 4 - *an **image** is a visual representation of an object, scene, person or abstraction, produced on a medium.*

As the focus of this thesis is primarily digital image collections, we need to present a definition of *digital images* as well. Digital images consist of many small dots, or *pixels*. Each horizontal line in the image has a fixed amount of pixels, as well as a fixed amount of such horizontal lines. For greyscale images, each pixel is represented by its brightness, or intensity. For colour images, each pixel is represented by three values representing one primary colour. Digital images can thus be represented using either a two-dimensional array, where each array element corresponds to a pixel (Greyscale), or three two-dimensional arrays, corresponding to the red, green, and

[3] For a thorough discourse into the nature of electromagnetic radiation and visible light, see for example Beichner, R. J. and R. A. Serway (2002). Physics For Scientists and Engineers, Thomson Leaning, Inc.

blue components of the image (Lu 1999). This leads naturally to the following definition of a digital image:

Definition 5 - *A **digital image** is a set of two-dimensional arrays composed of pixels whose locations hold digital colour and/or brightness information which, when viewed at a suitable distance, form an image.*

2.1.3 Concerning the Content of an Image

What is the content of an image? Depending on the viewpoint of the observer, image content can be described different levels of abstraction. In ordinary, everyday image use, the observer is usually interested in the objects present in an image, and the information and meaning that can be gained from these. In fields such as cultural studies or art history, the observer might be interested in the semantic content, or the stylistic and formal means used to create the image. In some technical disciplines, images are even regarded as a specific form of signal[4], where the important content is defined in the syntactical structure of an image, such as colour distribution or shapes present.

In order to have a clear understanding of image content, we need to examine the structural and conceptual components of an image. Jaimes and Chang (2002) presents a framework defining these basic concepts.

Percept vs. Concept
Images are representations of light, perceived by our visual senses. We can refer to this as the *percept* of the image. Patterns of light are reflected on different materials, and produce the perception of different elements such as texture and colour. This is the image *data*. The *concept* of an image refers to a representation, or an abstraction or generic idea generalized from the particular instances of the percept. As such, it implies the use of background knowledge and an inherent interpretation of what is perceived. This thus refers to the *information* that can be derived from the image.

Syntax vs. semantics
While percept refers to the impressions we perceive through our senses, the syntax refers to the visual elements themselves and the way in which they are arranged. For example, a colour blind person might perceive an image differently from someone with perfect vision, while the syntax of the image is the same. Semantics refer to the meanings of the syntactic elements and their arrangements.

Visual vs. non-visual content
An image's visual content corresponds to what is directly perceived when an image is observed, that is, which objects are seen on the image, such as shapes, colours, textures and so on. The non-visual content corresponds to information that is closely related to the image, but not present. The relationship between visual and non-visual content can be further illustrated by the classification of image content given by Panofsky (Woodrow 1999; Østbye and Schwebs 1999):

[4] A signal is an abstract element of information, or more specific usually a flow of information, in either one or several dimensions.

- The primary, or natural, content. Images are described based on the structures; the lines, colours and shapes in the image. Identification of image content is based on familiarity, such as persons, artefacts or landscapes. This type of identification is based on practical, day-to-day knowledge.
- The secondary, or conventional, content. Images are described, classified and interpreted by the motives, or symbols, present. This is based on cultural conventions, and requires knowledge of customs and traditions of both the culture and the time the image was created.
- The iconological, or intrinsic, content. Here the ideas, or the mind set, of the period when the image was created, are reflected. These are manifestations of the spirit of the times, or the "Zeitgeist", and a deep knowledge of these is necessary in order to understand, or read, the image.

Each of these levels requires a progressively higher and more specialized knowledge in order to "read" correctly.

Based on this, we can make an important distinction between the *syntactic* and the *semantic* content of an image:

Definition 6 - ***Syntactic Image Content*** *is the structure of an image; colour, texture, shapes and the spatial arrangements of these.*

Definition 7 - ***Semantic Image Content*** *is the meaning of an image, beyond its overt subject matter, including the emotional, intellectual, symbolic, thematic, and narrative connotations.*

Each of these can be subdivided further. Jaimes and Chang (2002) presents a 10 level structure, providing a systematic view of the different layers of image content, shown in Figure 3 below.

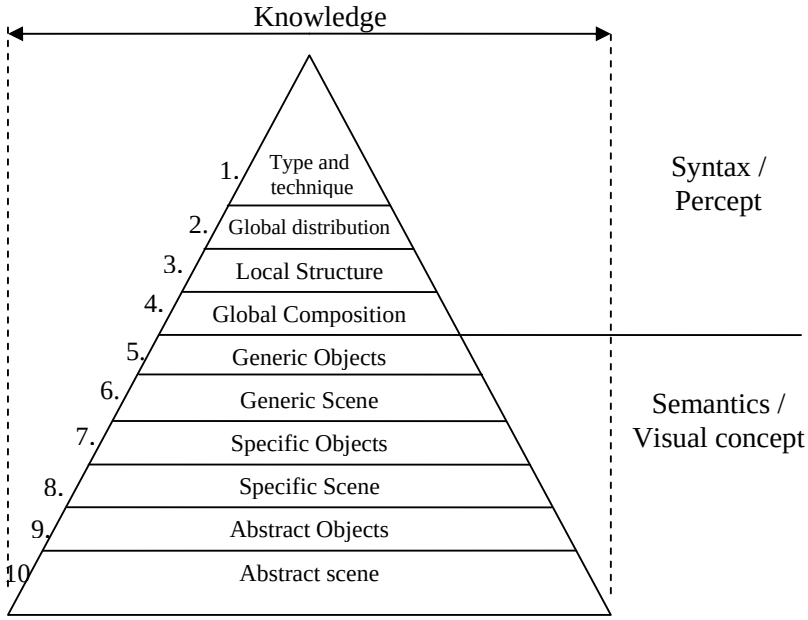


Figure 3 - Classification of image content, from Jaimes and Chang (2002).

The width of each layer of the pyramid represents the amount of knowledge required for recognizing and interpreting the image content.

Syntactic content

- Type and Technique are the general visual characteristics, such as format (Black and white or colour) or techniques used to create the image.
- Global Distribution refers to overall image characteristics, such as colour or texture distribution.
- Local structure is concerned with the different *components* of an image, as opposed to the global distributions. These are defined by shapes, colours and textures present.
- Global composition refers to the spatial structure of the image components, such as where they are placed in relation to each other.

Semantic content

- Generic objects refer to the *basic level categories*, where the basic concepts are defined, such as "dolphin" or "ball". This is the level where there are attributes common to all (or most) members of a category.
- Generic scenes refer to an entire image, representing one or more objects as a whole.
- Specific objects refer to objects which can be identified and named.
- Specific scenes are analogous to generic scenes, but with named objects
- Abstract objects – at this level, specialized or interpretive knowledge about what the objects *represent* is applied. This is analogous to the third level in Panofsky's classification (Woodrow 1999; Østbye and Schwebs 1999).

Furthermore, the content of an image is given and defined by the *image features*:

Definition 8 - *an **image feature** is a prominent or distinctive aspect, quality, or characteristic of the image.*

An image feature can be either *syntactic* or *semantic*:

Definition 9 – ***syntactic image features** are the low level structures of an image; colour, texture, shapes and spatial structure.*

Definition 10 – *a **semantic image feature** is a characteristic describing the generic, specific or abstract content of an image.*

Our understanding of the semantic content of an image is based upon using our knowledge to interpret the percepts, or visual cues, in an image. These visual cues are made up from the structural image content. The colours, shapes and other syntactic image features amalgamate into recognizable *visual objects*. These visual objects are the basic semantically recognizable unit in an image, and allow us to identify the objects depicted in an image:

Definition 11 - *A **visual object** is a set of syntactic image features combined into a semantically meaningful unit.*

2.2 Digital Image Collections – Use and Users

The twentieth century witnessed an unparalleled growth in availability, number and use of images. Images surround us everywhere, and their importance and usefulness is apparent in all walks of life. They play a crucial role in fields as diverse as entertainment, journalism, advertisement, education and medical care. Images were traditionally taken with optical cameras, developed in laboratories and stored in a physical location. Even though images could be digitized, electronic storage and usage of them were limited by low resolution, inefficient compression procedures, low bandwidth for transmission and storage capacity. The most widely used tools for information management, databases, also had limited support for images. Until recently, it has been impractical to use computers for the storage of large image collections.

The previous decade saw a rapid increase in the size of digital image collections. With the increase in computer capabilities, the ability to store more complex data types in databases, such as video, audio and images, has been drastically improved. This, combined with a similar development in bandwidth and transfer speeds, has made possible the development of very large digital image collections. This has been further fuelled by the rapid growth of the World Wide Web. A 1997 survey estimated the number of images available on the web to be in the range of 10 to 30 million (Eakins and Graham 1999).

However, the tools for describing and retrieving images from these collections have not been proportionately developed. Consider the case of a collection of images describing maritime life; marine animals and related activities. The images have been made available to the public through the internet, supporting all of the “standard” image description and retrieval techniques. Which expectations should, and could, we have to such an application?

In order to answer this, knowledge of user’s expectations are necessary - what kinds of queries are users likely to perform on this image database? Why do users seek images, what use do they make of them and how do they judge the utility of the images they retrieve?

When a potential user seeks access to a desired image from the collection, it might involve a search for images depicting specific types of objects or scenes, such as horses or dolphins, or pictures of their natural habitats. Additionally, they might want to find images evoking a certain mood, or images containing specific textures or patterns, such as zebra stripes. An image has different kinds of attributes and characteristics which can be used as the basis for queries, such as:

- Data about the author of the image, where and when it was taken and so on.
- The presence of a particular colour, texture, shape or spatial features. For a museum application, this might include searching for images with deep blue colours, indicating underwater still images.
- The presence or arrangement of specific types of objects, such as a flock of seagulls or a pack of horses.
- Depiction of a particular type of event, such as a pack of sharks hunting or a flock of seagulls feeding.

- The presence of named individual objects, locations or events, such as an image of the killer-whale “Keiko”.
- Subjective emotions one might associate with the image, such as happiness.

Each of the queries listed above represent a higher lever of abstraction than its predecessor, and each query is more difficult to answer without reference to some body of external knowledge. (Eakins 1996) suggest a three level query classification schema, separating queries into *primitive features* (corresponding to the *syntactical features)*, *derived features*, involving some degree of logical inference of the identity of the objects depicted in the image, and finally *abstract features*, involving a significant amount of high-level reasoning about the meaning and purpose of the objects or scenes depicted.

However, this classification does not capture the differences between *generic* and *specific* objects and scenes. A new classification scheme is proposed, by extending the query classification schema presented by Eakins (1996) with the *generic* and *specific* levels presented in Jaimes and Chang (2002). This is presented in Table 1, below.

Table 1 - classification of image retrieval query levels.

Level	Attributes	Description
0	Retrieval by external features	Creator, date taken, format
1	Syntactical features	Colour, texture, shape and spatial features
2	Generic features	Generic objects and scenes
3	Specific features	Specific objects and scenes
4	Abstract features	Activities, emotions and meanings.

An image retrieval system should support queries of all these levels in order to fully support the requirements of potential users. In order to support this, the retrieval system needs both classification and description for all these levels, as well as the tools and techniques required to perform retrieval based on this.

2.3 Image Storage - Description and Classification

Whenever large amount of data are stored, be it in a book, a database or a library, some sort of structure for classification and indexing is essential. Without it, search and retrieval of the data becomes a very difficult, if not impossible, task. Consider, for example, the task of finding a name from an unsorted list of names, or a single image from a collection of several thousand images. Without any sort of classification / index criteria, this will be a serial search from beginning to end. Several such techniques have been developed over the years, and some are presented below. In addition, new techniques have been, and need to be, developed for digital image collections.

2.3.1 Basic Storage Techniques – Fundamental Concepts

First of all, even small amounts of data need some coherent structure in which it is collected, or a database. Nordbotten (2004) defines a database as:

Definition 12 – *A **database** is a logically coherent collection of related data, representing some aspect of the real world that is designed, built, and populated with data for some purpose.*

From this definition, a database solely consists of the data items themselves. A system which provides structure and access to the collection is also necessary, and this is provided by a *database management system* (DBMS). This system must provide facilities for *storing and retrieval of data*, as well as a *query processing system*. Moreover, the system should provide additional functions such as transaction control and rights management. Finally, there must be metadata and indexes for structural support. Since it is assumed that the reader of this text has basic knowledge of database management systems in general, this will not be elaborated further than the following definition:

Definition 13 – *A **Database Management System, DBMS,** is a system providing 1) a schema for specification of the information content of the database, 2) a database engine that supports storage, access to and modification of the database, 3) a language for definition and manipulation of the database.*

As mentioned above, some sort of structure for classification of data is essential. An *index* is the most basic of these structures. It is basically a sorted list of important data, or information items, with a reference to where they are stored:

Definition 14 – *An **index** is a structure that serves to guide, point out, or otherwise facilitate reference of important subjects.*

Furthermore, it is sometimes necessary to record metadata, or *data about data*. This is data such as description of the organization of the data, the various data domains, and the relationships between them. There are three categories of metadata; *contextual*, *structural content* and *semantic content*. Contextual metadata are data that are external to the meaning of the document. For images, this can be data about the author, when it was taken and so on.

Structural content describes the structure of the document. For images, this represents the first four levels in the classification presented by Jaimes and Chang (2002), such as format or colour distribution (Figure 3, page 10).

Semantic metadata are data characterizing the subject matter, or the semantic content, of a document. For images, this can be data about the objects or activities depicted. This is represented by the six last levels presented in Jaimes and Chang (2002).

Both contextual and structural metadata describe data that are external to the *meaning* of an image. Baeza-Yates and Ribeiro-Neto (1999) calls this *descriptive* metadata. Based on this, the following definitions for metadata are proposed:

Definition 15 - *Metadata are data about data; data which provide information about, or documentation of, other data managed within an application or environment.*

Definition 16 – *Descriptive metadata* *are data that describe characteristics external to the meaning of a document.*

Definition 17 – *Semantic metadata* *are data that describe the semantic content of a document.*

Several different metadata schemes have been developed, both for traditional text documents and multimedia data. Two such schemes, the Dublin Core Initiative and MPEG-7, are presented in the following chapter.

2.3.2 Text Based Image Classification – Metadata and Keywords

This is based on using verbal language, or textual descriptors, for annotation and indexing of images. Usually, this designates either description through free text or keyword annotation. Text based classification has high expressive power. It can be used to describe almost the content levels described in Jaimes and Chang (2002), and it is in principle easily extensible to accommodate new concepts. Furthermore, we learn to express ourselves verbally from the very beginning of our lives. For most people, verbal language is a natural and the most developed form of communication.

Many image collections and picture libraries use keywords as their main form of retrieval, often using indexing schemes developed in-house, which reflect the special nature of their collections. Others use indexing schemes developed for libraries and archives (Eakins and Graham 1999). A well known example of an in-house indexing schema is the indexing of *Getty Images*, a collection of contemporary images. They have developed a thesaurus compromising over 10.000 keywords, divided into nine semantic groups. Index terms are assigned to whole images, important objects, and their setting. Retrieval software then allows users to search and refine queries at a wide range of levels, from broad, abstract terms such as “Freedom” to the more specific, e.g. “child pushing a swing” (Eakins and Graham 1999).

Metadata- and keyword annotation form the basis of both the Dublin Core Metadata Initiative and the MPEG-7 standards. Both are attempts to standardize descriptions of documents.

The Dublin Core Metadata Initiative is claimed to be “*a simple yet effective element set for describing a wide range of networked resources*.” (Hillman 2003: chapter 1.2). The basic set consists of 15 elements, each describing an important aspect of the resource, such as *Title, Creator, Format* and *Description*. Each of the DC elements is optional and repeatable, and there is no defined order. Furthermore, the elements can be extended by a set of *qualifiers*, such as *alternative* and *content*, used to further refine the elements *Title* and *Description*, respectively. Two broad categories of qualifiers are recognized. First, they can be *refinements* of a qualifier, making the meaning of an element narrower or broader, i.e. *Extent* or *Medium* for the *Format* element. Furthermore, they can be *encoding schemes*, aiding the interpretation of an element value, i.e. *URI* (Uniform Resource Identifier) for *Identifier* or *Source* (Hillman 2003).

The content of some of the elements can be selected from a *controlled vocabulary*, such as a thesaurus or a taxonomy. The use of a clearly defined set of terms can improve search results, since computers are good at matching words character by

character but weak at understanding the way people refer to one concept using different words, i.e. synonyms. Without basic terminology control, inconsistent or incorrect metadata can profoundly degrade the quality of search results. For example, without a controlled vocabulary, "candy" and "sweet" might be used to refer to the same concept. Controlled vocabularies may also reduce the likelihood of spelling errors when recording metadata.

While the Dublin Core dataset is relatively easy to implement and use due to its simplistic nature, it is unlikely that it is able to provide a rich enough indexing scheme for retrieval of images, especially when it comes to image content, as none of the elements support full descriptions for this. A given resource's content is described by the *Subject* and *Description* elements, which consist of both keywords and free text. These have to be manually entered, copied or automatically extracted from the item if there is no abstract or other structured description available (Hillman 2003).

MPEG-7 is an ISO/IEC standard developed by MPEG (Moving Picture Experts Group). It is formally named "Multimedia Content Description Interface", and "*provides a rich set of standardized tools to describe multimedia content"* (Martínez 2003). It offers a *framework* for describing content of multimedia objects, as well as metadata about form, conditions for access, classification, relationships with other material and context.

Although MPEG-7 allows for more complex descriptions than the DC metadata set, it is still just a *framework* for description. It is possible to create very complex definitions of image content, ranging over all the levels described by Jaimes and Chang (2002). While it is more complex than the Dublin Core, it is still relatively easy to implement. However, as with Dublin Core, MPEG-7 in itself does not provide any mechanisms for filling its elements, these have to be manually entered, copied or automatically extracted from the item.

We see that both Dublin Core and MPEG-7 offer good frameworks for describing and classification of images, but they do not address *how* the description and annotation is created. Traditionally, both descriptive and semantic metadata annotation of images created manually. However, depending on the equipment used for creating the image, as well as the system used to store the image, some of the descriptive metadata can be automatically generated by extracting data from an image. For example, some cameras now have the option to store data about when the image was taken, name of the photographer and similar contextual data, which can possibly be embedded in the image and extracted automatically. Furthermore, most syntactic image features can be automatically extracted using tools such as Oracle's *OrdImage* (Ward 2001; Hove 2003).

The problems of Volume and Subjectivity

No satisfactory solution has yet been found for automatic generation of *semantic metadata*, and these have to be entered manually. As long as image collections were small, this did not pose a significant problem. However, as image collections grew larger, manual annotation became prone to the problems of *volume* and *subjectivity*.

The problem of *volume* refers to the fact that manually annotation of an image is a time consuming task. Indexing times quoted in literature range from about 7 minutes per image from stock photographs at Getty Images, to more than 40 minutes pr image for a slide collection at Rensselaer Polytechnic (Eakins and Graham 1999). While it is relatively easy to create annotations for a small number of images, even a small personal computer now has the possibility to store millions of images, making manual annotation a daunting task, at best.

Furthermore, the combination of rich image content and differences in human perception makes it possible for two individuals to have very diverging interpretations of the same image. As a result, the description is prone to be both subjective and incomplete. Consider the image in Figure 4, below.

Figure 4 - A man feeding the killer whale 'Keiko'.

One possible textual annotation of this image could be “the killer whale ‘Keiko’ being fed by a man”. However, this does not include the name of the man, when and where the image was taken, or the context of the depicted situation. Furthermore, while some might see this as an image representing how humans and animals interact, others might regard this as an example on how animals are exploited by humans. This is called the problem of *subjectivity*.

The Problem of Explicability

Finally, while text based classification has a high expressive power, there are some limitations when dealing with visual objects. Some syntactical image features are difficult to describe with words. For example, although we have a set of terms describing the different colours, none of these terms are exact. Every colour has a broad range of different shades and intensities. Although most people are able to differentiate between two different shades, it is difficult to express the differences verbally without using fuzzy terms like “more” or “less” red. Furthermore, creating exact and objective textual descriptions of textures or shapes is difficult. We call this the *problem explicability*.

Combined, the problems of *volume, subjectivity and explicability* indicate that text based description is not sufficient for indexing and classification of images.

2.3.3 Syntactical Feature and Data Pattern Based Classification

A different approach to indexing and classification of images originates from the fields of Computer Vision and Pattern Recognition. This form for classification is based on analyzing the syntactic image features, and creating statistical and mathematical descriptors of these features.

The features are normally stored and represented by statistical and mathematical means, and each feature may have several such representations. *Feature extraction* is a term describing the process of analyzing an image, generating mathematical and statistical *descriptors* and using these or storing them for future use. The following definitions are proposed:

Definition 18 – *Feature extraction* *is the process of analyzing a digital image, used to generate mathematical and / or statistical descriptors of the syntactical image features.*

Definition 19 – *Feature descriptors* *are mathematical and / or statistical representations of syntactical image features.*

Usually, the descriptors are in the form of a vector, and are indeed commonly referred to as *feature vectors*:

Definition 20 – *A* ***feature vector*** *is a set of descriptors describing one, or more, syntactical image features, represented as numeric quantities.*

A feature vector is unique to a particular image, and is often called an image's unique *signature*, representing a "fingerprint" of the image.

Which descriptors are needed, or extracted, is usually dependent on the application area. Some applications might need very accurate descriptors of one or two syntactical features. For example, fingerprint identification needs very specialized descriptors of shape, and maybe texture, while more general applications need less specialized descriptors of all syntactical features. Feature extraction should meet the following requirements (Lu 1999:64):

1. [Descriptors] should be as complete as possible to represent the content of the information items.
2. The features should be represented and stored compactly. Complicated and large [descriptors] will defeat the purpose of feature extraction; it would be faster to search and compare information items themselves.
3. The computation of distance between [descriptors] should be efficient; otherwise the execution time would be too long[5].

Several different feature descriptors are used for the different image features. A brief presentation of some of these descriptors is given below. Further details concerning use of the descriptors are discussed in chapter 2.4.3. Shape descriptors have been given special attention, and are discussed separately in chapter 2.5.

Colour

Colour is an important dimension of human visual perception that allows discrimination and recognition of visual information. Correspondingly, colour features have been found to be effective for indexing and searching colour images in image collections. Generally, colour descriptors are relatively easy to extract and

[5] Similarity comparison and distance computation is discussed further in chapter 2.4.3.

match, thus well suited for content based queries. Typically, the specification of a colour descriptor requires fixing a colour space and determining its partitioning.

The most used colour descriptor is the *colour histogram*, which can be extracted from images in different ways. A colour histogram captures the distribution of colours within an image or an image region. It is basically a statistical quantification of the distribution of colour in a given image.

For a more thorough introduction to colour feature extraction, see for example Smith (2002).

Texture

Texture refers to visual patterns with properties of homogeneity that do not result from the presence of only a single colour or intensity. Pictures of water, grass, a bed of flowers and so on contain strong examples of image texture. Many natural and man made objects are distinguished by their texture. Examples of texture are tree barks, clouds, water, skin and fabrics. Such textured objects are difficult to describe in qualitative terms, let alone creating quantitative descriptions required for machine analysis. The observed texture often depends on the lightning condition, viewing angles and distance, may change over a period of time as in pictures of landscapes. Typical textural features include contrast, uniformity, coarseness, roughness, frequency, density and directionality. Texture features usually contain important information about the structural arrangement of surfaces and their relationship to the surrounding environment (Li and Kuo 2002; Manjunath and Ma 2002).

There are two basic classes of texture descriptors; statistical model based and transform based. The first approach explores the grey-level spatial dependence of textures, and then extracts meaningful statistics as texture descriptors. The second approach is based on psychological measures, uses numerical descriptors of coarseness, contrast, directionality, line-likeness, regularity and roughness. This corresponds to the characteristics of the human visual system, and therefore seems well suited to a general image retrieval application (Li and Kuo 2002).

For a more thorough introduction to texture feature extraction, see for example Manjunath and Ma (2002).

Spatial composition

Spatial composition refers to the structural relationships between components of the image. There are two classes of these relationships. The first class, containing topological relationships, captures the relations between element boundaries. The second class, containing orientation or directional relationships, captures the relative position of elements with respect to each other. Examples of topological relationships are “near to”, “within” or “adjacent to”. Examples of directional relationships are “in front of”, “on the left of” and “on top of”. Naturally, the spatial structure of an image is dependant on other features. In order to identify *where* a component is related to another component, the *component itself* needs to be identified, through colour, texture, shape or potentially other features.

For a more thorough introduction to spatial structure, see for example Li and Kuo (2002)

2.3.4 The Thesaurus

One of the techniques, or tools, used in description, classification and retrieval of information is the *Thesaurus.* The word ”Thesaurus” is derived from Greek and Latin words which mean a “Treasury”, and has been used for several centuries to mean a lexicon, or a treasury of words. The first modern usage of thesaurus dates back to 1852, when the first edition of “Thesaurus of English Words and Phrases” was published by Peter M. Roget. Its role in information retrieval began fully in the early 1950s, particularly through the work of H. P. Luhn.

The World Science Information System of UNESCO, UNISIST, presents a thesaurus as

A Thesaurus may be defined either in terms of its function or its structure. In terms of function, a thesaurus is a terminological control device used in translating from the natural language of documents, indexers or users into a more constrained “System Language” [...]. In terms of structure, a thesaurus is a controlled and dynamic vocabulary of semantically and generically related terms which covers a specific domain of knowledge (Foskett 1997).

Foskett also lists the major purposes of a thesaurus:

1. Give a map of a given field of knowledge, indicating how concepts and ideas are about concepts are related to each other.
2. Provide a standard vocabulary for a given subject field, ensuring that indexers are consistent when they are making index entries.
3. Provide a system of references between terms which ensures that only one term from a set of synonyms are used for indexing one concept.
4. Provide a guide for users of the system so that they choose from the correct term for a subject search.
5. Locate new concepts in a scheme of relationships with existing concepts in a way which makes sense to users of the system,
6. To provide classified hierarchies so that a search can be broadened or narrowed.
7. Provide standardized terms for a given subject field.

Based on this, (Baeza-Yates and Ribeiro-Neto 1999) defines a text based thesaurus as

Definition 21 – *A **thesaurus** is*

(1) a precompiled list of important words in a given domain of knowledge and
(2) *for each word in the list, a set of related words.*

Identified here are the two main parts of the Thesaurus; *index terms* (1) and *term relationships* (2). The *index terms* are the individual words, terms, or phrases. These are the basic semantic unit for conveying ideas. The terms are usually single-word nouns, since nouns are the most concrete part of speech. Adjectives and adverbs seldom convey any useful meaning, while verbs can be converted into nouns; *cleaning*, *reading* and so on. When terms are ambiguous, a “scope note” can be added to ensure consistency, give directions how to interpret the term. Naturally, not every term needs a scope note, but their presence is of considerable help in using a thesaurus correctly and reaching a correct understanding of the given field of knowledge.

The *term relationships* are links between the terms, often describing hierarchical relations, or synonyms and near-synonyms. Figure 5 below shows an illustration of thesaurus terms. Synonyms, and other useful relationships between terms, are called *related terms* (RT). The way the term *Cybernetics* is related to *computers* is an example of these relationships. Hierarchical relationships are used to narrow or broaden the scope of a term. *Broader Terms* (BT) generalizes a term, i.e. "Apparatus" is a generalization of "Computers". Reciprocally, *Narrower Terms* (NT) is a specialization, i.e. "Digital Computer" is a specialization of "Computer". BT and NT are reciprocals; a broader term necessarily implies at least one other term which is narrower.

Computers

Broader	Apparatus
Narrower	Analog Computers
	Digital Computers
Related	Automation
	Computer Application
	Cybernetics

Concept Learning

Narrower	Nonreversal Shift Learning
	Reversal Shift Learning
Related	Concepts
	Learning

Figure 5 - Excerpts from the *Thesaurus of Psychological Index Terms*. From Foskett (1977).

2.4 Image retrieval – Tools and Techniques

2.4.1 Fundamental Concepts of Image Retrieval

What is image retrieval? Different research fields have different perspectives and opinions on what an image is, and offer different theoretical foundations for, and approaches to, image retrieval. In this thesis, image retrieval is considered a part of the field of Information Retrieval (IR), a large and mature research area. Baeza-Yates and Ribeiro-Neto (1999:1) presents the field as dealing with "the representation, storage, organization of and access to information items". The purpose of information retrieval is to provide the user with easy access to the information items of interest.

Traditionally, the field of IR had textual documents as its main area of interest, and has been seen as a narrow area of interest mainly to librarians and information experts. However, with the development of information technology in the last two decades, the field has grown beyond this to accommodate other media types, such as images. The earliest techniques for image retrieval were based on techniques developed for text based information retrieval, but now also span techniques from such fields as artificial intelligence and signal processing. In order to have a clear view of what is to be understood with image retrieval in this thesis, the following definition is proposed:

Definition 22 – *image retrieval* *consists of the techniques used for query specification and retrieval of images from a digital image collection.*
As a research area, Image Retrieval is concerned with retrieval of images generated the whole electromagnetic range. However, this thesis is focused on retrieval of images captured using visual spectrum technologies. This is known as *Visible Image Retrieval* or *VIR*:

Definition 23 – *Visible Image Retrieval* *is the retrieval of images from heterogeneous collections of single images generated with visible spectrum technologies* (Colombo and Del Bimbo 2002).

In order to understand what challenges image retrieval systems face, we need to take a look at some of the existing techniques and their strengths and weaknesses. Consider again the scenario of a collection of images describing maritime life; marine animals and related activities. The images have been made available to the public through the internet, supporting all of the "standard" image description and retrieval techniques. Next, consider the case of a teacher preparing a lecture on dolphins; anatomy, habitat, feeding habits, activities and so on. The teacher wants to use images to illustrate his lecture, and approaches the aforementioned collection.

The teacher has certain *information needs*. He probably has an idea of the kind of images he wants to retrieve; what animals and other objects they should depict, how the objects are depicted, how these objects are arranged and so on. Furthermore, he might have need for images depicting dolphins from a specific angle, in a specific pose or involved in a certain activity. Next, the teacher has no idea of *which* images are contained in the collection, only the fact that it contains images of marine life, and maybe the number of images in the collection. The main problem facing the teacher is; how is he going to express his information need to the image collection and retrieve the desired images?

Image Retrieval techniques can be separated into two main groups; text-based image retrieval, and data pattern- and feature based image retrieval. A brief presentation of the two approaches as well as a list of the most used techniques is provided in the next chapters. While this is neither exhaustive nor in-depth, it presents an overview of the issues of image retrieval for the purposes of this thesis.

Dunckley (2003) presents two main approaches for image retrieval; *textual* and *visual*. Both approaches are used for both *query specification* and *image search and comparison*, resulting in four different categories, illustrated in Figure 6:

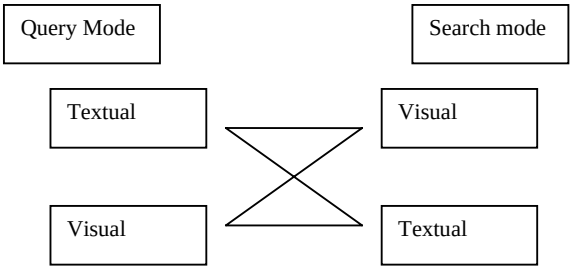


Figure 6 - Query and search modes for image retrieval systems. From Dunckley (2003).

Textual – Textual (TT) mode operates on the basis of forming a query with textual terms, then search through the textual image annotation in order to locate and retrieve images. Some text based query specification and image retrieval techniques are discussed in chapter 2.4.2.

Visual – Visual (VV) mode is based on comparing descriptors of syntactic image features in a query image to similar descriptors describing images in the image collection. Some visual based query specification and image retrieval techniques are discussed in chapter 2.4.3.

Visual – Textual (VT) and Text – Visual (TV) are based on the same query specification and retrieval techniques as VV and TT, but with a translation between query translation and retrieval. VT is based on using syntactic features to identify image content and use this to perform retrieval based on textual metadata. (TV) operate the other way, using textual retrieval techniques to identify visual image content, and use this content to perform visual query techniques for image retrieval. The mechanisms for translation between visual and textual representations is closely tied to a semantic interpretation of image content; unless it is possible to identify the semantic content in an image, it is at best very difficult to create a correct mapping between a textual and a visual term, and the other way around. This is one of the central motivations in this thesis project. The following two subchapters discuss the "pure" techniques (VV, TT). The combinatory approach is discussed further in chapter 2.5.

2.4.2 Text Based Image Retrieval Techniques

Text based image retrieval refers to retrieval based on the textual description of the contents of the image collection. Retrieval is based on similarity between a textual search string, and the keywords or free text annotation of the images in the database. The most used query types are keywords or text strings expressed by a user.

Text based queries allow all retrieval techniques from traditional, text based information retrieval to be used for image retrieval (Baeza-Yates and Ribeiro-Neto 1999; Lu 1999). Using this query type, it is in most cases easy for the enquirer to translate his information need into a query. And, as long as the images have been properly annotated, it is possible to achieve high levels of relevant image retrieval, for most of the query levels described in Table 1 (page 13).

First of all, we have image retrieval based on *external* image features (level 0 queries). These queries are based on the *contextual* and *structural* metadata describing images, such as format or size, photographer, copyright holder or date taken. Examples of such queries might be "Find images taken by PHOTOGRAPER before 2000" or "Retrieve all jpeg images smaller than 200x200 pixels". These are simple queries, which can be handled by DBMS capabilities, as long as the data exists in the database.

Next, we have queries based on the *syntactical* image features (level 1 queries). These are the most problematic queries for an image retrieval system based on text retrieval alone. Considering that images are of a visible nature, some information needs are difficult to express in text or keywords.

Consider, for example, either a search through a collection of trademarks or texture samples. In the first case, someone might want to use the image in Figure 7 below, as a logo for a company or a product line, and wishes to check if the logo infringes on existing trademarks.

Figure 7 - Illustration of a trademark.

While it is easy for a human observer to recognize shapes similar to the shape in Figure 7, it is at best very difficult to give a precise, objective and non-biased textual description of this shape. The same applies for colour, texture and other syntactic image features. Even though most people have a clear understanding of the colour "red", there are so many different nuances and shades that correspond to the word, and there might be different names depending on who you ask. Even if one manages to give a textual description of the shape, colour or texture, retrieval of similar objects using keywords or annotations is not guaranteed, since the results are dependent on the existing annotation.

Next, we have queries based on *semantic* image features (level 2 through 4 queries). These queries are based on the *semantic*metadata describing the content and meaning of the image, such as "Show me images of killer whales" or "Retrieve all images annotated with the keywords 'dolphin' and 'feeding'". Retrieval based on these queries are similar to retrieval based on level 0 queries, and are handled by text-based information retrieval techniques.

The main problem with text based image retrieval techniques comes from the problems of volume and subjectivity; the images in the collection must be correctly and completely annotated. Missing, or incorrect, image annotation will result in poor search results. No matter how good the query is formulated, it will fail if the annotation is inadequate. Furthermore, the enquirer does not necessarily share the vocabulary of the annotator, which might lead to a mismatch between the search criteria and the annotation.

2.4.3 Syntactical Feature and Data Pattern Based Image Retrieval Techniques

This approach to image retrieval is based on using syntactical feature descriptors as a basis for similarity searches between a query and images in the image collection. There are basically three approaches to query specification for this form for image retrieval; queries based directly on the syntactical features present in images, Query-by-Example and Query-by-Sketch.

Queries based on features, or data patterns, are based on a similarity comparison between a query set and existing patterns in the image collection. Information items are retrieved based on how similar descriptors or patterns of the image set are to the query patterns. An example of such a query could be "Retrieve all images with a

similar colour distribution to this colour histogram" or "Find images containing similar texture patterns to this set of textures". To answer this type of query, statistical information about database items should be pre-collected and stored.

These queries are useful when the enquirer is able to present a clear pattern to compare images, as in the trademark example above, as well as for other syntactical image features. However, in that the queries need to be statistical or mathematical in nature, it might be difficult for a non-expert user to express their information need in a query.

The second approach is based on *Query-By-Example* (QBE). QBE was originally the name of a domain calculus query language for relational databases, proposed by Zloof (1975). In image retrieval, use of QBE for query specification is one of the most used methods for visual query specification, and virtually all current CBIR systems now offer this form for searching (Eakins and Graham 1999). An example of this query would be "Find images similar to this image". These queries are useful when the enquirer has a clear vision of the images he wants to retrieve, or already has a relevant image to be used as a query. However, the use of QBE is somewhat problematic, as it relies on the enquirer having a suitable image at hand, and this might not always be the case.

The alternative approach is to accept user drawn sketches as basis for a query. This approach is known as Query-By-Sketch (QBS). Most of these approaches are based on the ideas originally developed for IBM's QBIC System (Flickner, Sawhney et al. 1995). QBS is based on letting the users sketch their own example images, either sketching the images by free hand or building seed images from images such as rectangles and circles. However, there are two problems with QBS. The actual tools and methods used to draw sketches might be simple and easy to use; most of us learned to draw before we could write. However, creating *good* drawings might be an entirely different matter. If the enquirer has limited artistic abilities, he might not be able to create a freehand drawing looking anything like the images he wishes to retrieve.

Queries based on data patterns, as well as QBE and QBS, are based on similarity between a query and the images in a collection. The similarity is calculated using *similarity functions* measuring similarity between syntactic feature descriptors. Even though there are a large number of widely differing similarity functions for computing all kinds of similarities, most of them is based on mapping pairs of feature vectors to a number, which is then representative of the similarity between two images. We have the following definition:

Definition 24 – *A **similarity function** is a mapping between pairs of feature vectors and a positive, real-valued number, which is chosen to be representative of the visual similarity between two images* (Li and Kuo 2002).

Usually, the number represents the Euclidean distance[6] between two feature vectors. If, for a given feature, two images are identical, the similarity function should be

[6] The Euclidean distance is the straight line distance between two points.

equal to 0. In other words, the less distance there are between images, the more similar they are.

So far, we have discussed image similarity on a global scale; syntactical features are extracted from, and compared to, whole images. For whole image matches, a single feature vector is extracted from each image and used for indexing and retrieval purposes. This framework was adopted in early CBIR systems, such as IBM's QBIC (Flickner, Sawhney et al. 1995) While this might be useful for comparing global image features, such as colour distribution, it is insufficient for identification and comparison of objects within an image. Furthermore, in addition to the interesting visual objects, an image is likely to contain some degree of noise:

Definition 25 – *Noise* *is irrelevant data that hamper the recognition and interpretation of the data of interest.*

Note that what might be regarded as interesting objects for one application area might be considered noise in another, and the other way around. In order to separate the interesting objects from the noise, there needs to be a mechanism for *segmenting* the image into subimages:

Definition 26 – *Segmentation* *is the process by which an image is divided into spatial sub regions.*

Segmentation can be either *data-dependent* or *data-independent*. Data-independent segmentation commonly consists of dividing an image into overlapping or non-overlapping fixed-size sliding rectangular regions of equal size and extracting and indexing a syntactical feature vector from each such region. This type of segmentation is easy and quick to perform, but generates a large amount of data. In addition, there is no guarantee that the segmentation is semantically meaningful. For some application areas, such as satellite imagery, this does not pose a problem, as one might expect large areas with similar texture. However, for images where there are few, important objects, data-independent segmentation is likely to divide the image in non-optimal locations, i.e. splitting a visual object over several regions.

Data-dependent segmentation is based on dividing the image based on its content, for example trying to identify objects, such as persons, from the background in photographic images. This type of segmentation produces fewer sub regions than data-independent extraction, and the ensuing segmentation can be used for automatic semantic labelling of image components. However, it requires more specialized tools and algorithms in order to produce semantically sound results. One example of this type of segmentation is *Blobworld* (Carson 1997), in which images are segmented simultaneously using colour and texture features. This method is well-tailored toward identifying objects in photographic images, providing they stand out from the background. A similar example of data-dependent segmentation is the neural network based algorithm presented by Rowley, Baluja et al.(1998). This algorithm is trained to identify and segment faces in photographic images.

A major difficulty facing syntactical feature and data pattern based retrieval techniques, is that an image is a two-dimensional representation of a three-dimensional space. The possible variances in scale, rotation and orientation of visual

objects are nearly unlimited. Consider the case of a tree. At a distance, it can be described as a blobby top attached to an elongated bottom. However, as one approaches the tree, large branches become visible, then smaller branches play a role, followed by leaves and so on (Kimia 2002). Furthermore, even very small changes in pitch, rotation or lightening conditions between two images of one object might lead to major changes in the syntactical image features. In order for similarity comparison based on data patterns alone to be effective, the depicted objects must have a high degree of visual invariance:

Definition 27 – *Visual invariance* *is the quality of an object to be resistant to variations in visual appearance.*

Consider the two images in Figure 8, below. Although the two images depict the same visual object and are very similar in semantic content (Both depict "a jumping dolphin"), they share few syntactical similarities. While this might not pose a problem for domain specific image retrieval with homogeneous images, we see that this is major challenge for visual image retrieval. Images, even depicting the same object, are often heterogeneous in nature, and retrieval techniques based on syntactic features are by default not capable of overcoming this problem for a general application area, as they lack understanding of semantic concepts.

Figure 8 – Two different images of a "Dolphin Jumping".

What the user wants from an image collection is a semantically meaningful answer to a query, and not only images who share a visual resemblance to a query. Research is needed towards developing tools and algorithms for identifying basic objects and semantic concepts in images are required.

2.4.4 Concerning the Difference between Textual and Visual Data

One of the axioms in image retrieval research is the claim that there is a structural difference between data presented in visual and textual form. And, because of this, it is at best difficult to use tools and techniques developed for textual data on visual data, and vice versa. A closer inquiry into this shows us that there might not be such a large difference between the two from a data management perspective.

It as been claimed that while text is *structured*, images are essentially *unstructured* (Eakins 1996; Santini and Jain 1997; Lu 1999). The claim is that textual material in electronic form has already been logically structured by the author in words, sentences and paragraphs using highly symbolic codes, such as the ASCII character set. For example, the character "A" can be represented as a byte, consisting of the bit string

“01000001”, which is easily interpreted by a computer. Words, sentences, paragraphs and entire documents can be broken down into these basic code units. Note that this does not allow a computer to have either a semantic *understanding* of the text, or interpret the text and extract information from it. However, it enables a computer to perform pattern matches between a query string and text in a document collection. For example, if the byte string “dolphin” is used as a seed for a text search, it is relatively easy for a retrieval system to compare this string to a text collection, find matches and return the documents containing the query term. If an information retrieval tool, such as a thesaurus is used, it allows retrieval of terms which are semantically related.

Images, on the other hand, contain a massive amount of data whose organization is spatial. The basic units of digital images are pixels, or arrays of pixel intensities, which have no inherent meaning. Even though a human is capable of identifying the content of an image through learned cognitive processes using sensory input, this process is much harder for a computer.

A comparison between the basic units of a text and an image will help illustrate this further. In the example above, 9 characters (7 letters surrounded by two “spaces”) identify the string as a unit, or a word. Now, consider the image in Figure 9, below. An uncompressed version of this image would normally consist of a large amount of pixel arrays, describing the colour and light intensity of each of the pixels. The claim is that there are no clearly predefined basic unites, such as ASCII code, present in this image, thus presenting a different, and more difficult, challenge than text based retrieval.

However, is it possible that there is an artificial separation between the “structured” data of text and the “unstructured data” present in an image, created by the fact that processing of images has been forced into the existing framework for text retrieval? Is there an actual data management difference between text and image retrieval? And because the techniques for identifying the basic units in an image are underdeveloped compared to text, is there any *real* difference between the two concepts?

First of all, use of the term “structured” for text is problematic. In terms of data management, “structured” data is *regular* or *tabular* data, where the semantic “meaning” of each row is known. Even though there usually is some structure to a text, such as titles and paragraph, a text is not tabular; it does not have the characteristics of structured data from a data management perspective.

Furthermore, it is postulated that an image does not have similar basic code units such as the alphabet. However, for an image to be interesting, it has to be an image *of* something. An image is never produced in a void. Even if an image only contains blackness, or even just random noise, it nevertheless contains *something*. Moreover, most images of interest have a certain composition; depicted objects are placed in relation to each other, or involved in specific activities, or they contain one specific object or scene. In some disciplines, such as media sciences, an image can be considered as a text, expressed in figurative language rather than verbal language.

Now, consider the image in Figure 9, below. If we regard the image as a collection of pixel arrays, the above assumptions on unstructured data hold. However, consider the image as a text in itself, with the objects depicted in the image as the basic code units.

Figure 9 - Image of a dolphin, a ball and two persons.

The image depicts a scene with a dolphin, two women and a ball, situated in what could appear to be theme park. The same scene could be described using a natural language, and translated into text, which again could be broken down into the characters making up the words and sentences. The verbal and the figurative text both tell a story. For most humans, reading the basic contents of an image is trivial. We learn to recognize ordinary shapes, persons, objects and even events in an image long before we can read. A human interpreter of an image is instantly able to identify the two shapes in Figure 9 as humans, and would be able to even if they were sitting, or involved in an entirely different activity.

Now consider Figure 10 below. Figure 10a presents a simplified model of an indexed text collection, while Figure 10b presents a similar model of an image collection. Both collections share several characteristics:

- Both collections are comprised of unstructured data
- The contents are stored as data types such as *long char* and *bit string*, represented as a string of bits
- Both are "interpreted" by a computer, either as characters or pixels
- Both are representations of semantic units; terms and objects
- Both can be enhanced with tools, such as a thesaurus, in order to improve search results.

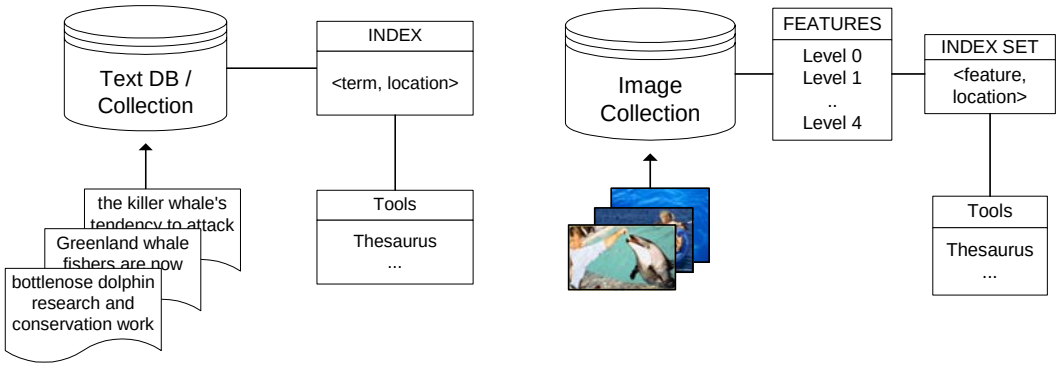


Figure 10 - Simplified models of a text document collection (a) and an image collection (b).

Is there a real data management difference on the conceptual level? There are many similarities, and it is believed that the distinction between textual and visual data might be artificial, and that by developing proper tools and techniques, it should be possible to treat the different texts similar.

The largest objection to this claim is the difference in invariance between the basic semantic objects. Consider again the case of the word “dolphin”. This is term is a textual representation of the concept of “a maritime animal of the Cetacean family”. There might be different terms describing the same concept in different dialects and languages, but for practical retrieval purposes there are a limited number of variations in the terms describing the concept – it has a reasonable degree of invariance. However, in terms of visual variance, there is virtually an endless amount of possible representations, and a low degree of invariance.

However, the difference in invariance between visual and textual data does not invalidate the claim that there is no *real* data management difference between the two data types. Rather, it indicates that better and more powerful tools are required for visual data. And if this is the case, it should be possible to use similar approaches for management of the two data types.

2.5 Image shape features

We have seen that shape is one of the important syntactical features describing visual objects and visual image content. However, understanding certain concepts is central to understanding the ideas and foundations on which this thesis is based: What is shape? How is shape represented, and how is it used in image retrieval systems?

The use of shape as a visual cue in image retrieval is less developed than the use of colour and texture, mainly because of the inherent complexity of representing it. Yet, retrieval by shape has the potential of being the most effective search technique in many application fields, as it has been shown that it is one of the main cues used by the human perception to interpret visual impressions (Sclaroff 1997; Kimia 2002).

A large body of literature exists on these topics, and even a brief overview of this lies outside the scope of this thesis. In this chapter, some fundamental concepts are presented, along with a presentation of one interesting approach to shape representation; *shape prototypes*.

2.5.1 What is Shape?

Let us begin our discussion with a general presentation of the characteristics of a shape, presented by Carlin (2000):

> *The shape of a physical object is the external form or contour, the geometry of its external surfaces or contours, the boundary between the objects interior and the exterior. Shape is the outline or characteristic surface configuration of the object.*

This broad presentation captures some of the most important characteristics of shape. First of all, *shape* is a geometrical term; it represents the spatial arrangement of an

object. The shape represents the contour of an object, or the boundary between the object and its environment. Furthermore, the shape of an object can be said to be invariant to variances in location, scale and rotation of the object; it represents the characteristic surface configuration of the object. This is reflected in the following mathematical definition of shape, given by (Dryden and Mardia 1998) :

Definition 28 - *The **shape** of an object is all the geometrical information that remains after location, scale and rotational effects are filtered out.*

This gives us a definition of what a shape *is*. However, we need two more important definitions before we can continue our discussion of image shape features. First of all, a *shape* is defined by its *contour*:

Definition 29 – *The **contour** of a two-dimensional shape is a continuous curve in the plane.*

The contour presents a complete representation of a two dimensional shape, and it can be specified by an infinitely large number of points in the plane.

Finally, the contour represents the border, or *boundary*, between a shape and its environment:

Definition 30 – *The **boundary** of a shape is the border between a shape and its environment, represented by the contour*.

With these definitions in place, we can give a closer examination to how shapes can be used in image retrieval; how can they be extracted and represented, how do we measure similarity between shapes?

2.5.2 Shape Identification, Extraction, Representation, and Similarity

We recall from our previous discussions that syntactical image features can be used to *index and classify* image (and image content), as well as being an aid in *image retrieval*. In order to use shape as a tool for image indexing and retrieval, shapes must be identified in, and extracted from an image. We have already seen how other syntactical image features can be extracted and used in image retrieval applications. Unfortunately, extraction and use of shape features are considered to be one of the most difficult aspects of image retrieval based on syntactical patters. The main reason for this is the difficulties of object segmentation and the variety of ways in which a three-dimensional object can be projected into two-dimensional shapes.

According to Li and Kuo (2002), there are two main steps involved in shape feature extraction; *object segmentation* and *shape representation.* Object segmentation is possibly the most challenging part of shape feature extraction. Generally, it is very difficult to perform precise, automatic object segmentation owing to the complexity of the individual object shape, the existence of noise and occlusion. There exists a large body of literature describing different techniques and approaches for object segmentation, such as region growing, edge detection and texture-based techniques. However, most of these techniques require a considerable amount of interactive guidance in order to achieve satisfactory results (Sclaroff and Liu 2001). One possible approach to automatic segmentation is to use a *model-based* approach. This makes use

of pre-defined models as an aid in identifying, and segmenting, shapes from the background. Shape prototypes, as detailed below, are one possible model-based approach for image segmentation.

Once objects are segmented, their shape features can be represented and indexed. In general, shape representations can be classified into three different categories; boundary based representations, region based representations and combinations of these (Li and Kuo 2002).

Boundary based representation emphasize the contour of the shape, and consists of techniques and algorithms used to detect and identify these contours. These techniques include *Fourier descriptors*, *Chain Code* and *Circular Arcs*. Interior based representation emphasizes the material within the boundary, or the "matter" of the shape, and techniques and algorithms for describing this. These techniques include *Moment Invariants, Zernike Moments, Morphological Descriptors and Deformable Shape Templates*.

Both boundary and region representations are complete, and can be used as a basis to compute the other by filling in the interior region or by tracing the boundary. Because both correspond to meaningful perceptual dimensions, an ideal representation would include both, enabling a full range of queries. Techniques for this include integration of different techniques, such as combining *Moment Invariants* with *Fourier Descriptors*.

A more thorough overview of the different shape segmentation and representation techniques can be found in (Dryden and Mardia 1998; Bouet, Khenchaf et al. 1999; Kimia 2002).

2.5.3 Shape Similarity Search

Once a shape has been extracted and a representation of that shape has been created, it can be used to measure similarity between that shape and other shapes. As with other syntactical feature comparison, shape similarity is measured using a *similarity function*. The actual functions used, and the inner workings of these functions, are intimately connected with the shape descriptors used.

Boundary based and region based representations inherently lead to different matching procedures. Boundary representations are typically accompanied by curve-based comparisons. In these approaches two curves are compared based on their properties, resulting in a single similarity measure.

Most of these techniques work along the same principles. A shape is usually described in a *shape feature vector*, containing a representation of a given shape. The similarity between two shape vectors can be computed as the Euclidean distance between them.

Region based representations typically involve trees, and have relied on such methods as *graph-tree matching, string edit distance* and *tree edit distance* (Kimia 2002). One such method, *deformable shape templates*, is detailed below.

Common for all techniques for shape similarity is that they share the need to adapt the matching process to the constraints of the shape representations. A more thorough overview of different shape similarity functions can be found in (Kimia 2002).

2.5.4 Deformable Shape Templates

The above discussion indicates that one of the main challenges facing the use of shape features for image retrieval is the difficulty of automatically detecting and segmenting shapes in an image. One possible solution is to use a *model based* approach to shape detection and identification. Model based approaches are based on exploiting prior knowledge, in the form of pre-defined models, as a tool during the identification and segmentation process.

One of the earliest, and most basic, model based approaches is to use simple *shape prototypes*, or shape templates:

Definition 31 – *A **shape prototype** is a static representation of a basic object or object category.*

These shape prototypes are basic representations of an object or object category for a given domain, such as tools, animals or household items. These prototypes, or templates, can be compared to an image in an attempt to identify shapes similar to the prototypes.

However, one problem with this approach is that the shape prototypes are only robust for very rigid shapes. Although many things have a high degree of invariance, in many cases this rigid-body model is inadequate. As noted in chapter 2.4.3, the possible variations in the shape of a dolphin are close to endless. Most biological objects are very flexible and articulate. To model these deformations, it is necessary to model the physics by which real objects deform (Sclaroff 1997; Sclaroff and Liu 2000).

One possible way to model these physics is *deformable* shape prototypes. Rather than directly comparing a candidate shape with all shape entries in the database, shapes are compared in terms of the types of non-rigid deformations that relate them to a small subset of representative prototypes. As with the shape prototype, each shape prototype is a representation of a basic object or object category. However, prototypes are allowed to elastically deform. The prototype is deformed to adjust to the shape of the query shape; the extent of the final match and the elastic deformation energy used in the deformation are used as a measure of shape similarity. Specifically, the contour of the prototype is allowed to deform into alignment with the candidate shape. We can then identify the minimal deformation between two shapes and use the size of the deformation to judge the similarity of the two objects (Sclaroff 1997; Carlin 2000).

Definition 32 – *A **deformable shape prototype** is a representation of a basic object or object category, capable of being deformed to align to a candidate shape.*

2.6 Image Retrieval Systems

2.6.1 Introduction and Basic Concepts

Techniques from traditional text based information retrieval alone have shown to be inadequate for successful image retrieval. When limitations of manual annotation became apparent, attempts were made to find other ways of indexing and retrieving images. Emerging from the field of computer vision, *Content Based Image Retrieval* (CBIR) was proposed as an alternative to manual indexing and retrieval.

The term CBIR seems to have originated in 1992, when it was used by T. Kato (Eakins and Graham 1999) to describe experiments into automatic retrieval of images from a database, based on the colours and shapes present. Since then, the term has been used to describe the process of retrieving desired images from a large collection on the basis of syntactical image features. The techniques, tools and algorithms that are used originate from fields such as statistics, pattern recognition, signal processing, and computer vision.

There is one problematic issue with the use of the term "Content Based Image Retrieval". We have earlier divided *image content* into *semantic* and *structural* content. The way the term CBIR is generally used, however, refers to the structural content only. This use excludes image retrieval based on annotation. Keywords and free text have the ability to give very rich and detailed description of image content. Even though this kind of image indexing and classification is prone the problems of volume, subjectivity and explicability, it can still be used to completely describe all levels of image content. It would be preferable if a narrower, yet descriptive, term could be used to describe syntactical feature extraction, as this would be more semantically correct. Nevertheless, CBIR has firmly established itself as a concept in image retrieval, and is used in this thesis. However, some definitions of CBIR limit the process to retrieval based on syntactic features alone. As was shown in chapter 2.1.3, the content of an image can be described at several different levels of abstraction. For the purposes of this thesis, I propose the following definition of CBIR:

Definition 32 – *Content Based Image Retrieval (CBIR)* *is the process of retrieving images from a collection based on automatically extracted features.*

One general problem with CBIR applications is that the more heterogeneous the application area is, the more difficult it is to develop good tools and algorithms. When dealing with homogeneous collections, it is possible to create algorithms tailored to that particular area. Consider the case of "MammoScan *μCaD*", a software tool developed to assist radiologists in identifying breast cancer in women (Engan, Fretheim et al. 2004). MammoScan *μCaD* analyses x-ray mammograms, trying to identify microcalcifications, one of the first signs of breast cancer. The images are very homogeneous, containing easily recognizable shapes without major differences between images. The program uses algorithms that are tailored to identify suspicious points on feature vectors extracted from the x-ray images. The algorithms "know" what the feature vectors are "supposed" to be, and what anomalies to look for. Similar situations exist in other homogeneous application areas, such as fingerprint identification. Although no fingerprints are completely identical, there are no large

differences in shape, structure or texture between different fingerprints, and it is possible to create specialized algorithms for similarity searches.

The above might allow us to conclude that the more specialized the image domain is, the easier it is to create good similarity functions and comparison algorithms.

Furthermore, some specialized image collections share a particular trait, differing from visual imagery – the fact that the *information lies in the structural components of the image*. Consider again the case of mammograms. The information which can be extracted from the image lies in the fact that anomalies, such as a tumour, are present in the structure. For fingerprints, the structure, or texture, of the image represents the information – the fingerprint is identifiable by its own structure.

Now, consider the image in Figure 11. This image has quite different information content than a fingerprint or an x-ray. Although it *might* be possible to extract some information from the structure, e.g. that the image contains a man and a dolphin, the image has also a higher level of semantic content, depending on how one interprets the image.

Figure 11 - A man feeding a dolphin?

For example, some might see this as an image of imprisoned animals, while others might see a dolphin playing with a man. It is at best very difficult to extract even some of this information with existing tools and techniques, as it is based on knowledge external to the image itself. One of the greatest challenges in image retrieval is to achieve higher levels of *semantic* image retrieval.

2.6.2 The Semantic Gap

Table 1 (page 13), presented five different levels of queries a potential enquirer might put to an image retrieval system. Each of these query levels represents a progressively higher semantic level. Furthermore, we have seen that while text based image classification and indexing is a powerful tool capable of describing most levels of image content, it is subject to the problems of volume, subjectivity and explicability, and thus not ideal for large scale image collections. Some sort of automatic image indexing is required to achieve efficient image retrieval. Syntactical feature and data pattern based indexing and retrieval techniques have been suggested as an alternative approach to image retrieval, and have proven to be useful and effective for retrieving images based on their syntactic features.

However, image retrieval based on syntactic features alone does not provide support for queries of level 2 – 4 (Table 1). There is a discrepancy between what the current solutions for this are capable of providing, and what we require of an image retrieval system. This gap between what is possible and what we require, has by many authors been dubbed *The Semantic Gap* (Eakins and Graham 1999).

Consider the four images depicted in Figure 12, below. Let us assume that the first image is used as a seed in a Query-By-Example search. A similarity search using syntactical features would first analyze the image and generate a feature vector describing the image. As it is a simple image, containing a clearly defined shape against a white background, we should expect that even the simplest algorithms would be able to identify the borders of the main shape. However, it would not have a semantic understanding of the concept "dolphin", and would not "know" what the identified shape is.

Figure 12 - Possible results from a query-by-example search.

Next, the descriptors generated from the seed image would be compared to the descriptors of the images in the collection using similarity functions. Depending on which syntactic features are compared, and the weights between them, we could expect the similarity function to retrieve the second image. Although the shape in this image is not as clearly outlined against the background as the seed image, a reasonably good shape-based algorithm should be able to pick up the shape. Although there are some differences between the two shapes, they are similar enough that the similarity function should identify the likeness between them and other images containing similar depictions of dolphins.

Taking a closer look at the third image, a human observer with a rudimentary knowledge of dolphins should immediately be able to recognize a dolphin in the image, or rather a dolphin's *beak*. However, if we focus solely on the syntactic features, we see that there are no obvious similarities between the two images. It might be possible that the similarity function would find likeness between the rounded head of the dolphin and the shape in the first image, or other unobvious similarities.

The last image is a black and white image of a banana. This has obviously *no* semantic connection to the concept of "dolphin". However, the syntactic features of the two images are fairly similar. If the two images exist in the same collection, it is more than possible that an unaided similarity function might retrieve this image[7].

The above example clearly illustrates the problems of current image retrieval systems based on syntactical features. Several problems can be identified from this. The feature descriptors and similarity function have no understanding the semantic features present in an image, even if it is capable of picking out a clearly defined object from the background. This has several implications:

- It makes combining text- and visual queries difficult. It is not possible to use the term "dolphin" as a query, and have the retrieval system retrieve images objects shaped as dolphins.

[7] The banana image was indeed retrieved in a test performed with Oracle 9i CBIR, using the first image as a seed.

- It makes it impossible to identify a dolphin from the image and retrieve images which have been (manually) annotated with the term “dolphin”.
- It is impossible for the retrieval system to retrieve dolphin images which are different with regards to the syntactical features.

Furthermore, even if the system *had* been able to identify the shape as a dolphin, and retrieve images containing dolphins in varying poses and angles, it lacks the ability to find images *related* to the concept of dolphin. Returning to Figure 12, the third image does not contain a whole dolphin, but a *dolphin beak*. Other images again might contain only other parts of a dolphin, such as a dolphin’s *fin* breaking the surface. Moreover, the dolphin is closely related to other animals, such as the killer whale, and these images might also be of interest to the enquirer.

Finally, lacking an understanding of the semantic content of an image, we saw that the retrieval engine might retrieve images with *no* semantic relationship to the seed image, such as a banana.

This presents us with this non-exhaustive list of problems:

- Non-retrieval of objects due to lack of object identification
- Non-retrieval of objects due to lack of relationship identification
- False retrieval of objects due to negative identification

Solutions able to alleviate these problems will help in narrowing the semantic gap. The first problem, lack of object identification, is of course the most important object and research needs to be focused towards this. However, it is possible that we might come closer to the problem of identification if we approach the problem from the angle of relationship identification. As a possible approach to this, we look to the field of text based image retrieval, and the thesaurus.

2.6.3 The Thesaurus in Image Retrieval Systems

We have seen that a thesaurus is a useful tool for improving information retrieval from text based sources by providing data about relationships between terms. One possible approach to identification of relationships between objects in visual images is to examine if the thesaurus framework can be adapted to this cause. However, let us first take a look at previous uses of the thesaurus in image retrieval.

Manjunath and Ma (2002) describe an architecture called a *texture thesaurus*. Using a self organizing map, a two-layered neural network, unique texture patterns are identified and classified. A hierarchical vector-quantization technique is the used to construct *texture code words*, each codeword representing a collection of texture patterns that are close to each other in the texture feature space. One can then use a visual representation of these code words as information samples to help users browse through the database. These collections, or code words, act as a template of the objects they represent. Manjunath and Ma illustrate this with an example of aerial photos. The thesaurus, and thus the code words, is based on landscape photos. For example, an image containing a parking lot is used as a seed image, and by comparing the seed image to the texture image thesaurus, the parking lot is identified and images containing similar code words are retrieved.

Dobie, Tansley et al. (1998) and Tansley, Bird et al (2000) describes a similar architecture in their multimedia information system MAVIS 2 – Multimedia Architecture for Video, Image and Sound. Central to the architecture is a *Multimedia Thesaurus*, which consists of *concepts* connected by *relationships*. Each concept is an abstract entity corresponding to a real world "object", and each concept is associated with one or more *media representations*, or multimedia objects that represent the concept. Media representations are associated with feature vectors, extracted from the representation using media processing algorithms. Different media representations of the same concept are considered equivalent and are called *synonyms*, even if they are of different media types, and they are linked to the same concept in the semantic layer.

MAVIS 2 has a four-layer data architecture, shown in Figure 13 below. The raw media layer consists of representations of all the new media objects that are usable within the system. A raw media object contains a reference to a media file, such as a web page or an image file, and some information about the type of the object. A selection layer contains selection objects that describe parts of raw media objects. For example, a selection may refer to an area of an image, an extent in a piece of text or a time segment in a piece of music. Furthermore, a Selection Expression Layer consists of selection expressions that describe combinations of selections and information about which properties of each selection are relevant. Finally, the *Conceptual Layer* consists of abstract representations of concepts, structured like a thesaurus. Each layer is related to the next, thus creating a chain of links between the raw media objects and the conceptual objects.

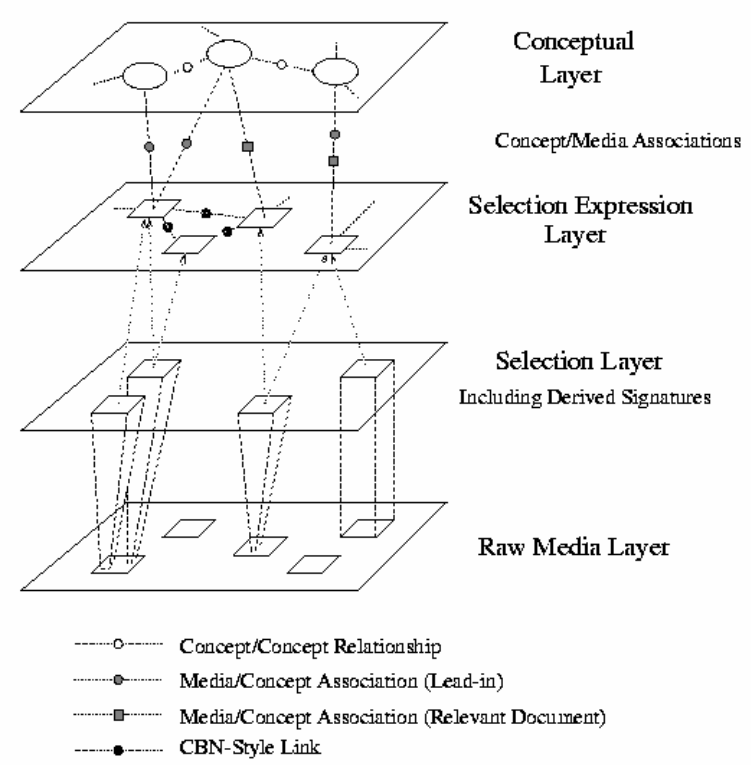


Figure 13 - Four-Layer MAVIS 2 Data Architecture. From (Dobie, Tansley et al. 1998).

The selection layer and the selection expression layer is similar to the *texture code words* in Manjunath and Ma (2002), in that these two layers act as representations of the objects in the conceptual layer, providing a link between the conceptual objects and the raw media data. These links can then be used to find images representing, or containing, the real world objects. Another similar approach is the use of *Visual Keywords* (Lim 1999; Lim 2000)

In other words, a thesaurus has been shown to be a useful tool for image retrieval, and an evaluation of a thesaurus for identifying and maintaining relationships between visual objects in images might prove an interesting pursuit. The framework suggested in chapter 3 is one possible approach towards this.

2.6.4 Architecture of Image Retrieval Systems

In chapters 2.3 and 2.4 some basic functionality of an image retrieval system is suggested. Such a system needs to have an underlying structure supporting storage, indexing, retrieval and manipulation of the collection. Huang and Rui (1999) suggests the following architecture for an Image Retrieval system illustrated in Figure 14:

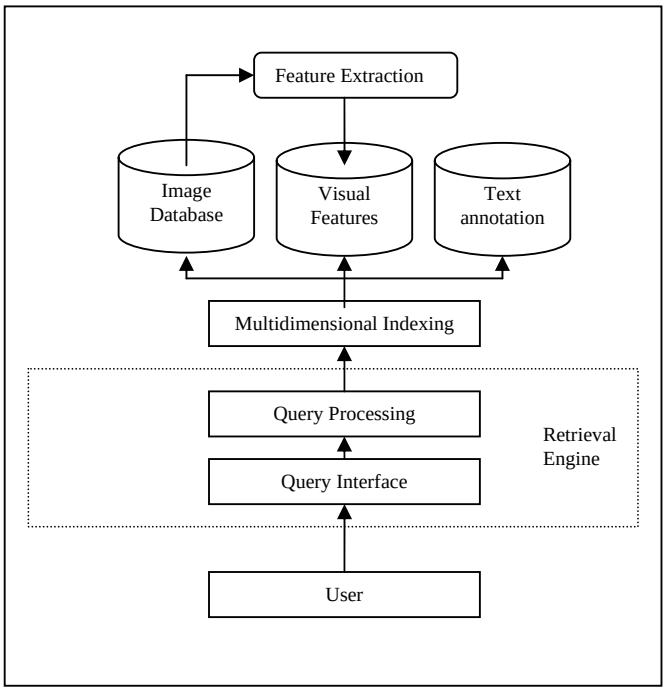


Figure 14 - Architecture of an Image Retrieval System. From (Huang and Rui 1999).

The architecture is based on four main components; a retrieval engine, an indexing structure, a set of data collections and mechanisms for syntactical feature based search and identification.

This architecture contains three data collections; the collection of raw image data for display purposes, a repository of visual features extracted from the images needed to support feature based similarity searches, and a text annotation repository containing keywords and free-text descriptions of images. There also needs to be some mechanism to extract syntactical features from the image and store them in the feature

repository. A multidimensional index supports fast retrieval as well as scalability to large collections. The retrieval engine includes a query interface and a query-processing unit. The query interface, typically employing graphical displays and direct manipulation techniques, collects information from users and displays retrieval results. The query processing unit is used to translate user queries into an internal form, which is then submitted to the system managing the collection, typically a DBMS.

While the illustration above might give the impression that an IR system should consist of three separate databases, this might not always be the case. On one hand, small systems might have all three collections integrated into one DBMS. Furthermore, larger systems might be distributed, consisting of several collections distributed on any number of servers and clients. Each of the data collections in the architecture above could therefore potentially consist of several databases, as mentioned in (Lu 1999). This is also reflected in main goal of the *Virtual Exhibits on Demand* project;

> *To develop methods and tools for searching multiple, multimedia museum databases for information that can be used by educators and/or students for construction of local presentations* (Nordbotten 2002).

However, it is felt that the architecture described above gives a good conceptual overview of the basic components of an image retrieval system.

3 The Shape Thesaurus

We have seen that there are problematic issues with the existing techniques for visible image retrieval. New approaches are required to reduce the Semantic Gap between currently available retrieval systems and the functionality we require. Motivated by this, a novel approach to image retrieval is suggested. Techniques from data- and feature based image retrieval are combined with the structure presented by a thesaurus. The approach is based on similar uses of thesauri in image retrieval, as described in chapter 2.6.3. The resulting framework, a *Shape Thesaurus*, is suggested as a tool which might help alleviate some of the problems related to the semantic gap.

3.1 Motivation behind the Shape Thesaurus

Two of the major problems with currently available syntactical feature- and data pattern based image retrieval techniques were identified as non-retrieval of correct images due to:

- Lack of correct object identification
- Lack of understanding of object relationships

From traditional text-based information retrieval, we know that a *thesaurus* is tool that can provide semantic relationships between different terms. For retrieval of textual data this allows retrieval of texts that are semantically related to the query.

The objective of this research project is to evaluate if a structure based on, or similar to, a text thesaurus is capable of improving image retrieval. Particularly, is it possible to use a similar structure to create and maintain links between shapes that are different in appearance but similar in semantic content? Furthermore, can such a structure also be useful for identification of visual objects? If this can be achieved, we will be one step closer to bridging the semantic gap. This presents us with the two main motivations for proposing a thesaurus for shapes:

- Assist an image retrieval system in identifying shapes that are representations of the same *semantic* content but share few, or no, syntactical similarities.
- Assist an image retrieval system in retrieving images containing shapes that are semantically *related* to the semantic content of a shape.

Based on this, we see that the main focus of the shape thesaurus is mainly focused at improving image retrieval queries of level 2 (Generic features) and, to some degree, level 3 (Specific features), described in Table 1, page 13. Level 0 queries (External features) will not be affected by a Shape Thesaurus and level 1 queries are focused at the syntactic, and not the semantic image features. It is possible that a shape thesaurus *might* be able to improve retrieval from level 4 queries (Abstract features), in that it will assist in identifying objects. However, these queries are mainly focused at the meaning, and the meaning that can be derived from the content, they are not the main focus of the shape thesaurus.

3.2 The Shape Thesaurus Defined

A shape based thesaurus can be defined as:

Definition 33 – *a **shape thesaurus*** is

> *(1) A precompiled list of important shapes representing visual objects in a given domain of knowledge*
> *(2) feature descriptors describing these shapes*
> *(3) a textual / semantic description of these shapes*
> *(4) for each shape, a set of related shapes.*

This definition identifies the four major components of a shape thesaurus.

3.2.1 Shape List

The shape list is a set of shapes representing important visual objects in a given domain. Each shape prototype is a possible representation of a visual object, and the visual objects are described by a set of shape prototypes. These shape prototype sets are equivalent of the "term list" in a text-based thesaurus. The shape prototypes chosen should provide the retrieval system with enough data to be able to recognize and identify image content.

The shape prototypes used should be representative of the important visual objects within the thesaurus' domain. The prototypes are used to identify the visual objects, and should be a representation of the visual characteristics of these objects. For the maritime scenario described earlier, the thesaurus could be based on a taxonomy of marine animals, and extended with objects that are important for visual identification of the visual objects. These extensions were shapes describing characteristic features of the thesaurus shapes, such as a "shark's fin" or a "dolphin's head". These are visual cues that let us recognize the animal they describe even if it is partly occluded.

3.2.2 Shape Feature Descriptors

While the first component provides a list of the important shapes, each of these shapes must be described by a set of feature descriptors. These descriptors are the most important facets of the shape thesaurus, as proper identification of the thesaurus terms depend on being able to find similarities between image content and the thesaurus shapes.

Choosing viable descriptors for representing shape, along with the algorithms for comparing shapes, is likely the most important decision for building the shape thesaurus. More specialized descriptors and retrieval functions have higher success in identifying shapes than generalized descriptors, and this alone suggests using highly specialized descriptors and retrieval functions when building an image retrieval system. However, as the descriptors and algorithms grow more specialized, the domain of the thesaurus grows narrower. This suggests that the choice of which descriptors should be used in an implementation of a shape thesaurus should be based on the characteristics of the domain of the application.

However, unless developing a shape thesaurus for a very specific domain, it is likely that the differences in shape between the thesaurus terms are large enough that very

specialized descriptors and similarity functions, such as developed for fingerprint matching, would be too restrictive. One possible solution to this could be to use different descriptors and algorithms for different subsets of the thesaurus domain. However, there are two distinct drawbacks to this approach. First, one would need a different set of descriptors and similarity functions for the different sub domains, depending on how specialized the domains are. It might prove costly and time consuming to develop and / or implement a large number of different descriptors / algorithms. In addition, the system needs to have some sort of mechanism to identify the correct sub domain for any given query. While the former drawback is mainly an economic issue, the latter is more difficult to overcome. In order for the system to select a suitable set of descriptors and similarity functions, it has to have some data about which sub domain the image belongs to, which basically leads us back to the original problem; how to identify the content of the image.

Based on the above, descriptors and functions capable of similarity matching on general terms would be preferable. Although it needs to be given a thorough study, it is believed that *deformable shape templates* could be a viable approach when choosing shape descriptors; each distinct thesaurus shape can be described by a deformable shape template This approach is general enough to allow description of different types of objects, and still use the same descriptors and similarity functions to the entire domain.

If two or more domains are combined to a single thesaurus, it could be possible to use different subsets of templates, i.e. one for marine life and one for vehicles. Note that using different sets of templates might lead to the same problems as described in the last paragraph; that is, the system might have difficulties deciding which subset to choose. A thorough study is needed before we can be more conclusive about which descriptors should be used.

3.2.3 Semantic Label

If a shape thesaurus should be made available directly to a user, there must be a textual / semantic description of the objects. Although there *might* be such a description of the shapes in the shape list, i.e. by giving semantic meaningful names to the shapes, it is probable that some of the terms might be difficult to understand and the descriptors are unlikely to convey any information to most users. It is therefore necessary to add a semantic label to each of the terms.

3.2.4 Thesaurus Relationships

Finally, we have the set of relationships between the defined objects. A text based thesaurus usually has 3 types of relationships; *Broader Term, Narrower Term* and *Related Term*. A shape thesaurus expands on this with two sets of relationships, as shown in Table 2, below.

The first set, the *object relationships*, is related to the visual objects the shapes represent. These relationships provide support for retrieval of images with content semantically related to the query items requested, but not identical. These relationships are akin to the relationships found in a text based thesaurus, in that they allow for both query expansion and query refinement. For example, a user querying after whales might want to narrow his search to a certain whale species, or expand his search to include whaling vessels.

Table 2 - Shape Thesaurus Relationships.

Relationship	Description	Example
Object relationships		
Generalization	An object is a generalization of its related object.	"Whale" is a generalization of "Blue Whale".
Specialization	An object is a specialization of its related object.	"dolphin fin" is a specialization of "fin"
Related	An object has an unspecified relationship to its related object	A "Whaling ship" is related to a "whale"
Shape Relationships		
Part-Of	A shape is a part of its related shape	A "Dolphin fin" is a part-of a "Dolphin"
Variant-of	A shape is a variant of its related shape.	A "jumping dolphin" is a variant-of a "swimming dolphin"

The second set, or the *shape relationships,* is related to the shapes representing the visual objects. These relationships provide support for retrieval of images representing different variants or sub segments of the same visual object. This addresses the problem with lack of visual invariance. First, the *part-of* relationship allows for retrieval of images where the object in question is only partly visible in an image. Consider the example in Figure 15, below. The only visible object of interest in this image, is the head (or beak) of a dolphin. The shape representing this object is only a *part-of* the shape representing a whole dolphin. By utilizing this knowledge, it is possible to retrieve images which contain *whole* dolphin shapes.

Figure 15 - A Dolphin's Beak.

Furthermore, the *variant-of* relationship allows for retrieval of images containing different shapes which represent the same object. Consider the images in Figure 16. Both images represent a single dolphin, but the structural differences are so large that any single algorithm comparing them would probably find no similarities between the two. Even so, they are semantically similar, in that they are both sharks. However, as was pointed out above, both images are very descriptive and familiar depictions of

dolphins. This type of relationship between shapes can be represented by a *"variant-of"* relationship.

Figure 16 – Illustration of visual invariance in different "dolphin" depictions.

Furthermore, the *part-of* and *variant-of* relationships can be used to identify the visual objects present in a shape. By matching a shape in an image to the shapes in thesaurus, such as finding high degrees of similarity between an area in the image and descriptors describing a dolphin, this might suggest that the shape is a *variant-of* a "dolphin". The retrieval system can then use this to retrieve relevant images.

Note that while the shape thesaurus provides support for the functionality described here, it is completely reliant on having good and effective similarity functions. If the similarity functions are unable to process an image, identify dominant shapes and compare this to the thesaurus shape descriptors, the shape thesaurus will not provide any useful support for the image retrieval system. This stresses the importance of good descriptors and similarity functions.

3.3 The Shape Thesaurus exemplified

An example of a potential structure of a shape thesaurus is shown in Figure 18[8] below. While this doesn't show the actual shape descriptors, it acts as a conceptual illustration of which visual objects can be included.

Variant-of relationships are difficult to represent in a hierarchical structure such as Figure 18, as they represent relationships between the *shape descriptors* representing variants of the visual objects, and are primarily visual in nature. Figure 17, below, presents a visualization of some of the shape descriptors used in the shape thesaurus implemented in the *VORTEX* prototype[9]. They represent three different variants of the visual object "Dolphin Beak".

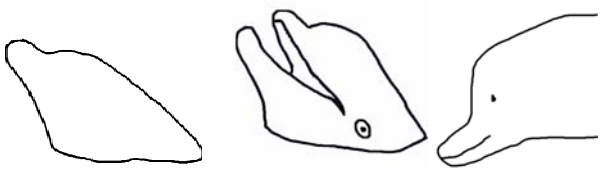


Figure 17 - A visualization of the variant-of relationships for the shapes representing a "Dolphin Beak".

[8] The structure of the thesaurus was based in part on the biological taxonomies found at http://www.nmnh.si.edu/msw/ and http://www.itis.usda.gov/index.html

[9] The shapes were created from actual images of dolphin beaks.

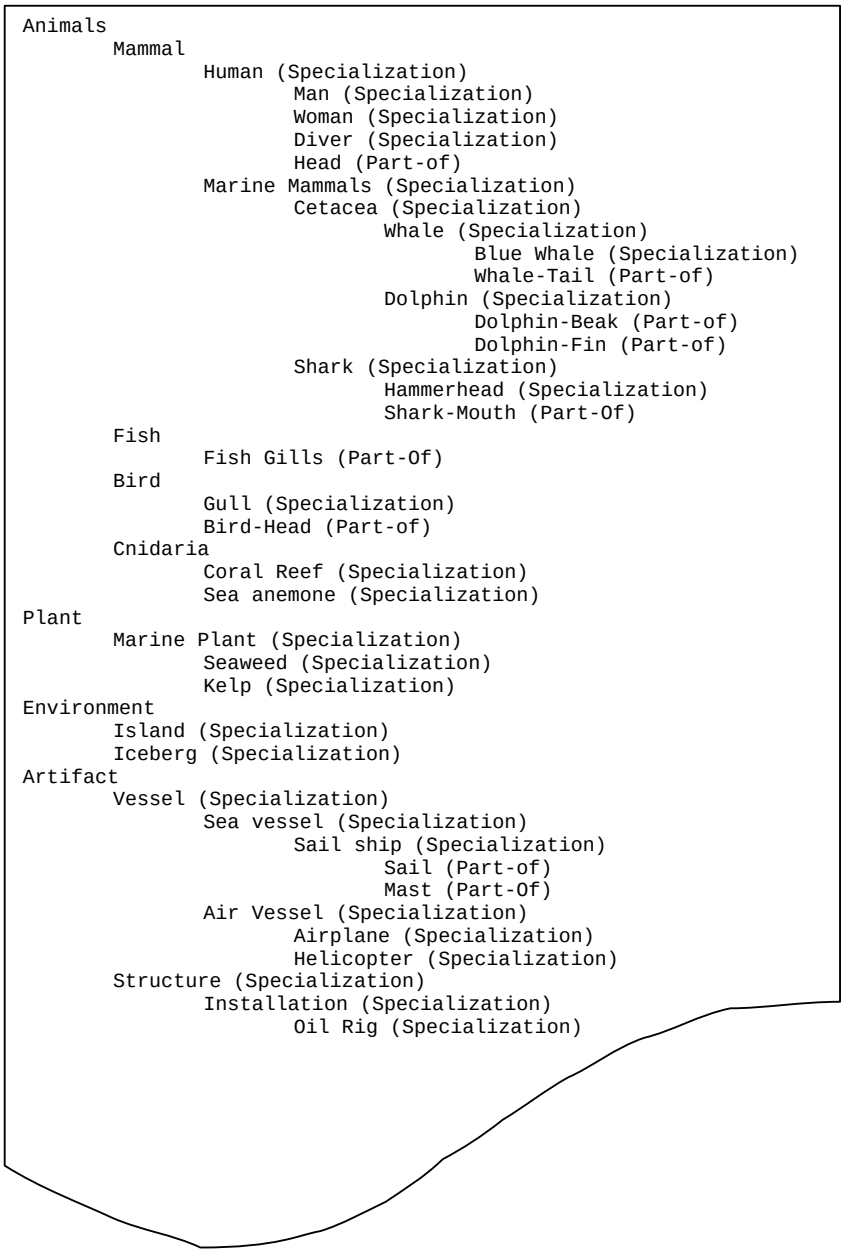


Figure 18 - Illustration of the objects and relationships in a Shape Thesaurus.

3.4 Using the Shape Thesaurus

We recall from Figure 6 (page 22) that there were 4 different modes for query specification for image retrieval. The shape thesaurus introduced in this thesis is intended as a tool for supporting three of these:

- Visual to visual
- Visual to text
- Text to visual

Figure 19 presents the different stages of the image retrieval process, showing which steps in the process shape thesaurus can provide support:

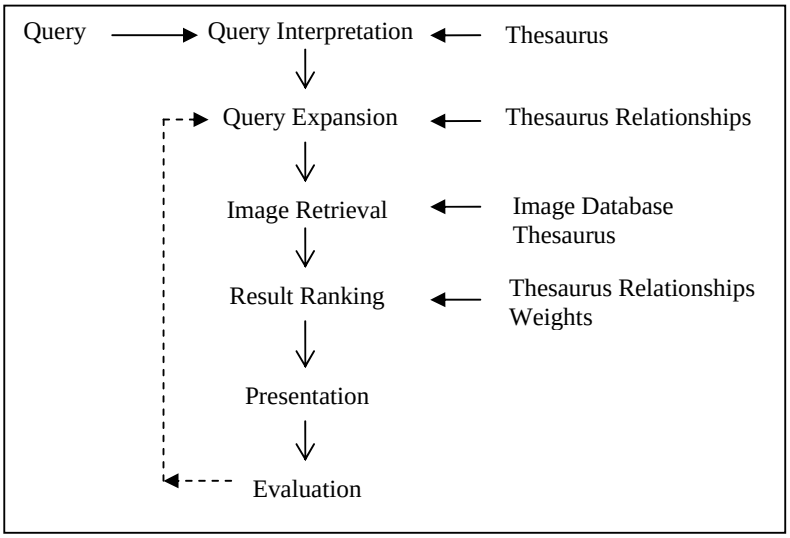


Figure 19 - Image Retrieval with a Shape Thesaurus.

Although the actual interface for the query specification, as well as presentation and evaluation, are integral parts of the image retrieval system, they lie outside the boundaries of the shape thesaurus, and are not discussed further here.

3.4.1 Query Interpretation – Identification of concepts

A shape thesaurus should be able to accept queries in both textual and visual form. In the case of textual queries, the query interpretation process is simply a matter of text based information retrieval techniques. The textual input is compared to the semantic labelling of thesaurus objects, and if a match is found the system can base the consecutive steps on this data. Additionally, the textual query can be used by any other, text-based retrieval tools present in the retrieval system.

However, in the case of visual queries (Query-By Example and Query-By-Sketch), the query interpretation becomes more complex. In order for the shape thesaurus to prove useful there must be a mechanism for comparing the query image to the shape descriptors. The success or failure of the framework rests on the ability to analyze images and identify the existence of thesaurus objects. If the system fails to make positive identifications of components present in an image, it will not be able to utilize

the thesaurus. Furthermore, and possibly even more serious, if the system makes a false identification, images with *no* relevance to the query will be retrieved.

One of the purposes of using a shape thesaurus is to identify the visual objects present in an image. In the case of visual queries, the seed images might have one, or several, objects present. Furthermore, the image is likely to contain some elements of noise. Although the user might be encouraged to use images containing clearly defined example objects, it is unlikely this will be the case for all seed images. Good segmentation algorithms are needed to separate and identify the visual objects in the image. The resulting subimages, or segmented shapes, should be used as a basic for comparison with the thesaurus descriptors.

The actual similarity functions available for comparing the identified shapes to the thesaurus collection are dependent on the shape descriptors used in the shape thesaurus. As such, the actual similarity functions should be chosen from state of the art functions available for the actual descriptors in the application domain.

3.4.2 Query Expansion

Once the visual objects in the query have been identified, the thesaurus relationships can be utilized for query expansion. First of all, the *part-of* relationship should be used to identify both objects containing the query object, as well as objects it is a part of. For example, if a “Dolphin Fin” was identified, images containing “Dolphins” and “Dolphin Beaks” could be retrieved. We see that “Dolphin” is directly related to “Dolphin Fin” through a part-of relationship. However, “Dolphin Beak” is only related to “Dolphin Fin” through the relationship with “Dolphin”. This indicates that there needs to be some specification about how many levels of relationships should be traversed. This could either be set as a default parameter, or specified by the user before the query or during user-specified query expansion.

Furthermore, the *object-relationships* could be used to identify objects that are semantically related to the identified term. Greater care should be used here, as more objects will result in a larger result set. As with the *part-of* relationship, specification of how many hierarchical levels are retrieved could be set as a parameter. It is likely that the object-relationships would be utilized most during user-specified query expansion, to broaden or narrow the search result.

3.4.3 Image Retrieval

During the image retrieval phase, the thesaurus can be used to retrieve images containing the objects identified during query interpretation. Three different methods for utilizing a shape thesaurus are suggested here.

First of all, if the images in the collection are annotated with text, it is possible to use the semantic labelling of the identified objects as a basis for text-based similarity searches. It is likely that images annotated with words similar to the semantic labelling will be relevant for the original query. The actual details of this search are dependant on the text-retrieval capabilities of the image retrieval system.

Furthermore, in order to fully utilize the thesaurus, there needs to be some sort of association, or mapping, between the thesaurus terms and the image collection. Without an association the usefulness of a shape thesaurus is limited to text based

searches, as described above. Two different associations between the shape thesaurus and an image collection are suggested here; *shape similarity* search and *direct links*.

First of all, it is possible to use the same similarity functions used to identify shapes present in an image to retrieve images containing similar shapes. This approach is based on using similarity functions to compare the descriptors describing the thesaurus objects to descriptors describing the images in the collection. The major strength of this approach is that it is able to find images without needing any image pre-processing other than generating and indexing descriptors.

The other approach is to create links between the thesaurus and the image collection. Objects in the thesaurus can be linked directly to the images containing depictions of them. These links can then be used to quickly retrieve images containing these objects. Three different techniques for creating such links are discussed here. First, the links could be created manually, either for images existing in the database or when new images are inserted. A human observer is, so far, much better at judging the content of an image than an automated process. This approach is likely to provide high degrees of correct and complete object identification.

However, manual pre-processing of the images is a time consuming task. In cases with large, unprocessed image collections, or when a large amount of images are inserted into an existing collection, manual linking becomes prone to the problems of volume and subjectivity. This suggests that manual linking is only applicable for small image collections, or when a small number of images are inserted into the collection.

Next, the links could be created automatically by the shape similarity functions described above. Automatic linking would not be susceptible to the problems of volume and subjectivity, and an automated process would be able to handle a much larger amount of images in a shorter period of time. However, automatic shape recognition is susceptible to false identification, resulting in incorrect links.

Finally, a semi-automatic linking process could be imagined. This method is based on interaction between image retrieval system and its users to create links based on a relevance / feedback structure. Users provide feedback as to whether a certain image is relevant to, or contains, a thesaurus object. While this is subject to some of the problems of manual linking, it would be possible to use a weighted average of the feedback from all users, thus correcting for subjectivity and individual errors.

However, the users have to be presented with an initial set of images, suggesting that there must exist at least some mapping between the thesaurus shapes and the image collection. This suggests that this method should be used in conjunction with either or both of the abovementioned methods. This could also be used to correct erroneous links, as described below. Choosing between manual and automatic linking should be balanced between the need for precise and correct links, and the time and effort required by manual linking.

A combination of the above approaches is likely to be a good approach for implementing a shape thesaurus framework. On one hand, it is unlikely that all images will be completely analyzed and all content successfully identified, unless the

collection is small. This again suggests that a similarity search is necessary. Furthermore, having established relationships between some images and the terms will give a faster, and perhaps better, search result than a similarity search alone. A summary of the different associations is presented in Table 3:

Table 3 - Methods for mapping thesaurus objects to an image collection.

Method	Strengths	Drawbacks
Direct link		
Manual linking	• High rate of correct identification	• Prone to problems of subjectivity and volume
Automatic linking	• Fast • Efficient • Not prone to subjectivity or volume	• Dependent on quality of the similarity algorithms • Prone to false identification and erroneous links
Semiautomatic linking	• Combines the strengths of both automatic and manual linking	• Might be subject to the problems of volume, subjectivity and false identification • Requires some initial mapping
Similarity search	• Not reliant on existing links	• Dependant on quality of the retrieval algorithms.

3.4.4 Result Ranking

After the system has retrieved the images found to be relevant, images in the result set should be ranked in order of relevance to the image query. Result set ranking assisted by a shape thesaurus can be based on the perceived *semantic closeness* between the query and the retrieved images.

It is assumed that the enquirer would find images containing the visual objects identified from the original query as most relevant, and should be presented first. Next, depending on how broad or narrow the search was defined, or refined, to be, images containing objects *related* to the identified objects. It would be natural to present *Part-Of* relationships first, as these are basically representations of the query objects. Furthermore, images containing objects *hierarchically* related should be presented in order of how many levels they are removed from the query objects. Finally, images containing objects which have a *general* relationship to the retrieved concepts could be presented. These are presented last as they usually represent looser relationships. One possible result set ranking is shown in Figure 20, below. Each parenthesis represents a higher level of semantic closeness, with the first and innermost images containing the identified objects. The actual order of the levels suggested here might not be optimal, and should be given further study. It is likely that the order should be partly based on user preferences; some users are likely to be more interested in images containing variant-of depictions, while others might want to have images containing *part-of* depictions first.

Resultset (((((Images with query object) Variant-Of) Part-Of) BT/ NT) Related)

Figure 20 – Example result ranking with a shape thesaurus.

4 The VORTEX Prototype

In order to evaluate the proposed shape thesaurus framework in an experiment, one possible implementation of an image retrieval system with a shape thesaurus has been prototyped. The prototype has been named VORTEX engine, Visual Object Retrieval – Thesaurus EXtension. The functionality of the prototype has been kept to the absolute minimum required for purposes of evaluation.

A software development project is a complicated process, and there exist different methodologies that define a framework for this process. The *Unified Software Development Process* (Jacobson, Booch et al. 1998) was used as a methodological framework and aid during the implementation of the VORTEX prototype. The Unified Process is a very thorough and descriptive process, and was not followed strictly during implementation of VORTEX, but rather adapted to suit this development project. The development process is documented through UML (Unified Modelling Language) diagrams, available in appendix B. UML diagrams are used visualization aids to describe the functionality of the prototype components, and a basic understanding of these diagrams is assumed.

4.1 Planning the VORTEX System

The framework described in chapter 3 presents the Shape Thesaurus as a tool which can be adapted and used in both existing and new image retrieval systems. For the purposes of evaluating the framework, a basic image retrieval system with a thesaurus extension (VORTEX) was developed from scratch. This was done in order to have as much control over the test environment as possible. The VORTEX system consists of four components; an image collection, a shape thesaurus, a basic CBIR retrieval algorithm and a retrieval algorithm based on the shape thesaurus.

4.1.1 Requirement Specification

The first step towards implementing the VORTEX prototype was to describe the required functionality of the four main components. The required functionality is described in Figure 21.

Based on the requirement specification, important *Use Cases* were identified. An overview of the use-cases is presented in Table 4, below.

Finally, based on the requirements and the use cases, an initial *UML analysis diagram* was created. This diagram, shown in Figure 22, shows the planned core components in the system.

The VORTEX system is an image retrieval system extended with a shape thesaurus, consisting of four main components; an image collection, a shape thesaurus, a basic CBIR retrieval algorithm and a shape-thesaurus aided image retrieval algorithm. The components should support the following requirements:

1. Image collection
   a. Storage space for digital images
   b. Functionality to add new images and generate syntactic feature descriptors

2. Shape Thesaurus
   a. List of important shapes for a given domain
   b. Semantically meaningful names of these shapes
   c. Syntactic feature descriptors of these shapes
   d. For each shape, a set of related shapes
   e. Functionality for administering the Shape Thesaurus
      i. Add thesaurus shapes
      ii. Add shape descriptors
      iii. Add thesaurus relationships
   f. Functionality for creating relationships between the thesaurus shapes and the image collection
      i. Manual update of relationships

3. Basic CBIR retrieval algorithm
   a. Compare the syntactic features of a seed image to the syntactic feature descriptors of the image collection, and return the images which are perceived as similar to the seed image

4. Shape thesaurus aided image retrieval algorithm
   a. Identify objects in a seed image using similarity functions based on comparison between the shape descriptors and the image descriptors.
   b. Provide support for query expansion through the relationships defined in the shape thesaurus
   c. Retrieve images which are perceived as relevant to the identified objects in a seed image
      i. Retrieve images using established links
      ii. Retrieve images using similarity functions

Figure 21 - Requirement specification for the VORTEX system.

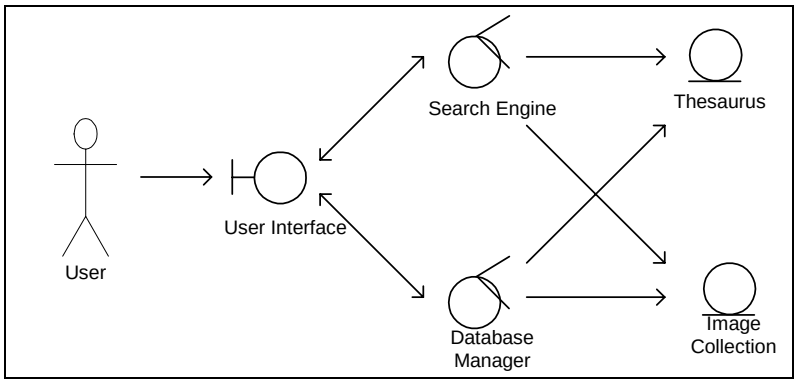


Figure 22 - UML Analysis Diagram.

Table 4 - Use Cases identified for the VORTEX prototype.

#	Use-Case Name	Event	Description
1	Add Thesaurus Object	This use-case describes the event that an admin adds a new thesaurus object.	The user supplies data about the new thesaurus object – shape name and semantic description. The system stores this data in the thesaurus.
2	Add Thesaurus Representation	This use-case describes the event that an admin adds a new descriptor for a given thesaurus term.	The user supplies an image representing a thesaurus shape to the system. The system analyzes the image, generates a signature and retrieves the image properties. The image and the extracted features are stored in the thesaurus.
3	Add Thesaurus Relationship	This use-case describes the event that an admin adds a new relationship betweens two thesaurus objects.	The user selects two thesaurus representations and specifies the relationship between them. The system will then generate two entries in the thesaurus, describing both sides of the relationship.
4	Search using standard CBIR algorithm	This use-case describes the event that a user searches for an image using a standard CBIR algorithm.	The user supplies a seed image. The system will analyze the image, compare it to the images in the image collection and return a set of images with high similarity scores.
5	Search - Thesaurus enhanced QBE	This use-case describes the event that a user searches for an image using the thesaurus.	The user supplies a seed image. The system will analyze the image, compare it to the term descriptors and try to identify thesaurus term(s) in the image. Next, the system will find a set of terms relating to the identified term. The identified term as well as the set of terms will then be used to find other images containing this set of terms. Finally, the system will compare the seed image to the entire collection and try to find similar images the thesaurus might have missed. The resulting images will be returned to the user.
6	Generate Link - Manual	This use-case describes that a link is created manually between an image and a thesaurus object	The user creates selects an image and adds a link between the image and an object in the thesaurus.
7	Add Image	This use-case describes the event that a user adds an image to the image collection.	The user supplies information about how and where the image is stored. The system will generate a signature, retrieve the image's properties and store the image, the signature and the image's properties in the database.

4.1.2 Development Platform and Software

Oracle 9i (O9i) toolkit was chosen as the primary development platform for the prototype. O9i is an object-relational database management system with support for both SQL/3 and object-oriented software development through PL/SQL[10]. Furthermore, with the interMedia toolkit, O9i offers basic support for image management. It includes syntactic feature descriptors and similarity functions through the *OrdImage* and *OrdImageSignature* classes. Finally, O9i supports the use of Java Classes, allowing for advanced functionality through existing or user defined classes.

However, the use of existing solutions clearly limits the ability to customize the syntactic descriptors and similarity functions. However, it allows rapid prototype development. Furthermore, as O9i provides native support for image storage and retrieval, little time and effort would be needed to fit the shape thesaurus prototype to the image database. Finally, implementing basic content based image retrieval functionality in O9i is a relatively straightforward task. This would allow for rapid development of a CBIR system used as a comparison to the Shape Thesaurus framework.

Furthermore, the existing software and database prototypes in the *Virtual Exhibits on Demand* project were created with O9i as a development platform. It was intended that the work presented in this thesis would be included in these prototypes. The process of adapting this framework to the existing VED prototypes would be easier if developed on the same platform. Finally, O9i was a familiar tool for me, removing the time and effort needed to become proficient with another development tool.

4.2 Implementing the VORTEX System

The VORTEX prototype was implemented as an object / relational database structure. The structure for the image collection and the thesaurus were created as PL/SQL objects and tables, shown in Figure 23, below. The model was created using the *Structural Semantic Model* (SSM) described by Nordbotten (2004).

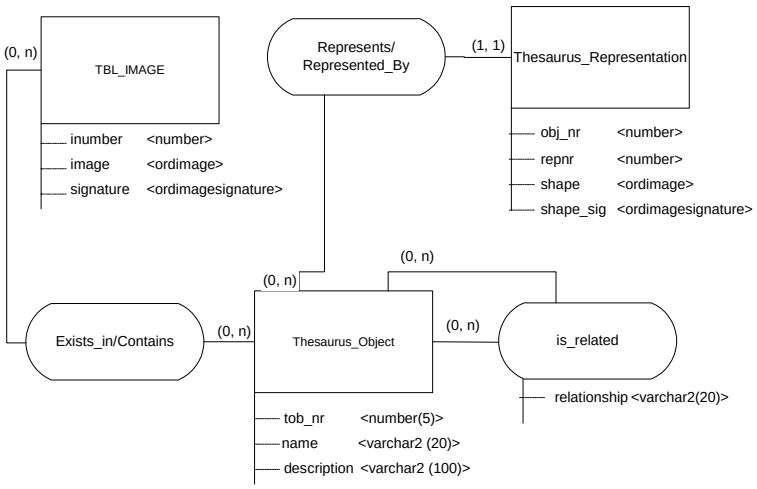


Figure 23 - SSM model of the VORTEX prototype.

[10] Procedural Language / SQL

Functionality for the image collection, thesaurus management and the two search algorithms were created using PL/SQL objects, methods and function. Figure 24, below, presents an overview of these. The *DBManager* object contains the functionality to import images; SearchEngine contains the actual search algorithms, while the ThHandler object contains functionality for maintaining and using the shape thesaurus. *ImageList*, *SignatureList* and *ThobList* are arrays used by the other objects for exchanging images, signatures and thesaurus objects.

A complete overview of all the different objects and their methods are given in appendix C. Important functionality and structures are highlighted in the following subchapters.

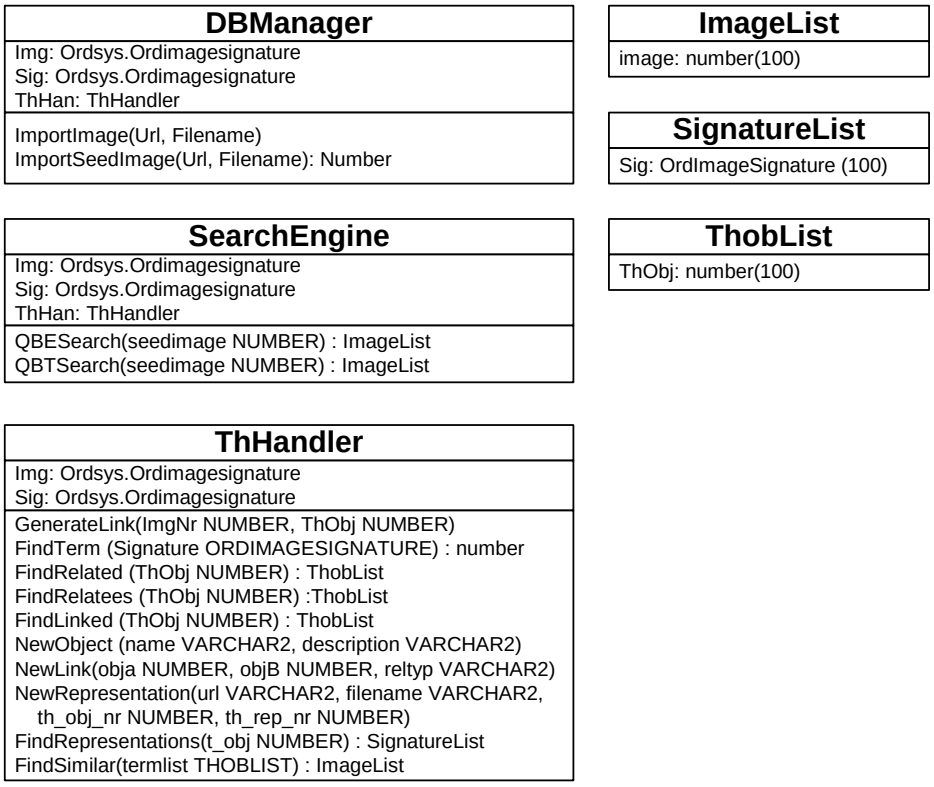


Figure 24 - VORTEX Class diagram.

4.2.1 The Shape Thesaurus

The actual structure of the shape thesaurus in the VORTEX prototype is defined by the *Thesaurus_object* and *Thesaurus_representation* entities, as well as the *is_related* relationship in Figure 23, above. Furthermore, the *ThHandler* class, shown in Figure 24, contains the functionality concerned with creating and maintaining the thesaurus, as well as the actual use of it.

Object list, Semantic Labels and Shape Relationships

The framework presented in chapter 3 separated between the actual shapes and the *semantic labels* describing these objects. However, the possibility of giving semantic

meaningful names to the shapes was proposed as an alternative. For the purposes of the shape thesaurus in the VORTEX prototype, the latter approach was used.

Chapter 3.3 presented an illustration of a possible structure and visual objects for a shape thesaurus based on a maritime scenario. However, only a subset of this was created for the shape thesaurus in the VORTEX prototype. Using a small subset would reduce both the effort required to build the thesaurus structure and the shape representations. If it could be shown that the framework proves useful for a small subset, it would be possible to generalize this to a larger set of thesaurus objects.

A short list of visual objects was compiled based on different animals and characteristic features of these animals. A schematic overview of the identified objects is presented in Figure 25. Hierarchical relationships are identified by the postfix "*Subtype*", with the term to the left and above the denoted term being the *generalization* and the denoted term the *specialization*. The *Part-Of* shape-relationship is identified by the *Part-Of* postfix. The denoted object has a *part-of* relationship to the above object.

Figure 25 - Visual objects in the Thesaurus.

In the shape thesaurus in the VORTEX prototype, the *variant-of* relationship type is created by having a set of different shape descriptors for each thesaurus object.

Note that the objects *animal*, *mammal* and *marine mammal* are included only for purposes of hierarchical relationships. They are not actual visual objects.

The object list has been implemented as a relational structure, shown as the entity *Thesaurus_Object* in Figure 23, above. Figure 26 shows a view of the *tbl_thesaurus_object* table, showing the table structure and the actual data.

TOB_NR	NAME	DESCRIPTION
1	Animal	Supertype for the animal kingdom
2	Mammal	Supertype for class mammal
3	Human	Superclass Human
6	Whale	The marine mammal Whale
7	Whale Tail	The characteristic tail of the whale

Figure 26 - Tbl_Thesaurus_Object showing labels and description of the thesaurus objects.

New thesaurus objects can be added to the thesaurus using the *NewObject* method of the *thesaurus handler* object. This is a simple function that generates a new row in the *tbl_thesaurus_object* table, adding a new object with name and description, but without any shape representations.

The object relationships have also been implemented as a relational structure, shown as the relationship *is-related* in Figure 23, above. Figure 27, below, shows a view of the actual implementation of this relationship, as well as some of the actual data.

T_OB_A	T_OB_B	RELATIONSHIP_TYPE
1	2	Supertype
1	3	Supertype
5	11	Supertype
9	8	Part-Of
10	8	Part-Of

Figure 27 - Transcript from TblTob_Related, showing the object relationships.

New thesaurus relationships can be added to using the *GenerateLink* method of the *thesaurus handler* object. As with the *NewObject* method, this is a simple function that generates a link by inserting a new row in the *tblTob_related* table.

Shape Templates and Shape Feature Descriptors

Selecting suitable shape-descriptors for the thesaurus objects was identified as one of the most important as well as difficult aspects of creating a Shape Thesaurus. This belief was fortified during the implementation of the shape thesaurus in *VORTEX* prototype. Generating shape feature descriptors is a two phase process:

1. Identify and generate the different shapes to use as object descriptors.
2. Generating shape descriptors of these shapes.

In the VORTEX prototype, the shapes were described by *shape prototypes* and *example images*. The shape prototypes are simple shapes representing one possible variant of a thesaurus object. Each thesaurus object is then represented by a set of these shape templates. In addition to being a representation of the visual object, the set of different shape templates represents the *variant-of* relationship. Each different template is a variant of the object they describe.

The shape templates were based on actual images depicting the visual objects. Considerable effort was spent in order to find images and shapes which could represent the objects they described. The shapes were manually extracted from images

using tools available in standard image processing software[11]. Each of the templates were then dimensioned to 400 by 400 pixels, and presented as either black lines or silhouettes on a white background.

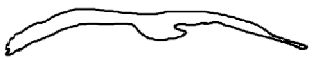
Figure 28 - Generating a shape template.

Figure 28 shows an illustration of the steps taken to generate a shape template. An image containing a "good" visual representation of a thesaurus object was selected. The image was trimmed to the actual object. The result of this trimming is shown in the first image. Next, visual noise was removed using either the built-in feature detection algorithms the photo editing tools, or by manual processing, resulting in the second image. For some images, such as shown in this example, very little manual effort was needed to isolate the visual object. Thereafter, the shape boundary was drawn out manually. In some cases, clearly defined structures, such as the dolphin's mouth and eye in the above figure, were outlined as well. Finally, the rest of the original image was removed, leaving only the outline of the shape.

A total of 77 different shape templates were created for the thesaurus objects, with an average of 6 shapes for each object. No shape representations were given for *animal* or *mammal*.

In addition to the shape templates, *example images* were used as shape descriptors. These are images containing good visual representations of the thesaurus objects. It was believed that these might help the similarity algorithms to identify the terms in an image. A total of 82 different example images were used, with an average of 7 images for each object. Some example templates and images are shown in Table 5, below. A complete overview of the shape descriptors can be found in Appendix D.

Table 5 – Examples of shapes used as thesaurus object descriptors.

Term	Template	Template	Example
Bird			
Diver			
Whale			

[11] Paint Shop Pro 8.0 was used as the primary image processing toolkit.

For actual feature representation and description of the shape templates and example images, the feature descriptors provided by O9i were used. O9i supports this extraction and storage of syntactical features through the ORDImageSignature class.

Oracle Corporation does not divulge the internal structures of the image signatures or their similarity functions, as they are considered business secrets. However, a brief presentation of the techniques used is presented by Guros (2004)

> *In the current implementation color is characterized using the HSV color model, and each computed distance between two values is weighted by the difference between the two colors in the HSV color space. Texture is characterized by contrast, coarseness and directionality. Shape is characterized by area, perimeter, circularity, aspect ratio and moment variants.*

There are some problematic issues with using such proprietary tools, mainly that the actual implementation and application of these is hidden from view. This leaves little room for customizing and adapting them to the purposes of the shape thesaurus. The consequence of this is examined further in chapter 6.2.2.

The shape templates and example images are stored in an object-relational structure, shown as *Thesaurus Representation* in Figure 23. Each shape thesaurus object is possibly related to many shape representations. The actual images are stored as *Binary Large Objects* in the *OrdImage* data type. The feature descriptors are stored as a binary string in the *ORDImageSignature* data type. Figure 29 presents a view of the object-relational structure used.

```
Name           Null?          Type
=======================================================
OBJ_NR         NOT NULL       NUMBER
REPNR          NOT NULL       NUMBER
SHAPE                         ORDSYS.ORDIMAGE
SHAPE_SIG                     ORDSYS.ORDIMAGESIGNATURE
```

Figure 29 - Structure of tbl_thesaurus_rep.

New shape representations can be added to the thesaurus using the *NewRepresentation* of the *ThHandler class*. The signature, or feature vector describing the shape (and other syntactical features), is generated by the *GenerateSignature* method of the *OrdImageSignature* when a new image is inserted.

Figure 31 shows the actual code used to import a shape and generate a feature vector. This method belongs to the *ThHandler* class. The procedure imports an image from a URL source, analyzes the image, and stores the image, as well as a syntactical feature vector, as *OrdImage* and *OrdImageSignature* data types in the *tbl_thesaurus_rep* table.

Figure 30, below, shows the code used to import a new thesaurus representation. This actual function inserts the first representation for the third thesaurus object, “human”, as denoted by “3, 1” in the fourth line.

```
declare
th thhandler := thhandler(null, null);
begin
th.newrepresentation('www.student.uib.no/~st04839/treps', ‘human_example_1.gif',3,1);
end;
```

Figure 30 - Importing a new thesaurus representation.

```
member procedure newrepresentation(url varchar2, filename varchar2, th_obj_nr
number, th_rep_nr number) IS

-------------------------------------------------------------------------------
-- This procedure adds a representation image to a  thesaurus object. Needs
-- the number of the thesaurus object, as well as the number of the
-- representation. This should be automated. In addition, supply the URL
-- and the filename for the image.
-------------------------------------------------------------------------------

t_image ordsys.ordimage;
t_image_sig ordsys.ordimagesignature;
ctx RAW (4000) := null;

begin

insert into tbl_thesaurus_rep
values (th_obj_nr, th_rep_nr, ordsys.ordimage.init(),
ordsys.ordimagesignature.init());

select shape, shape_sig into t_image, t_image_sig
from tbl_thesaurus_rep where obj_nr = th_obj_nr and repnr = th_rep_nr
for update;

t_image.setsource('http', url, filename);
t_image.import(ctx);
t_image_sig.GenerateSignature(t_image);
t_image.setProperties;

update tbl_thesaurus_rep set shape = t_image, shape_sig = t_image_sig
where obj_nr = th_obj_nr and repnr = th_rep_nr;
end newrepresentation;
```

Figure 31 - The *NewRepresentation* function, used to import new thesaurus shape representations. The actual signature generation is highlighted.

4.2.2 Image Database

The image database was kept as simple as possible while still maintaining the functionality required to work together with the Shape Thesaurus. It simply consists of one object/relational structure, shown as *TBL_IMAGE* in Figure 23. The feature descriptors and the actual images are generated and stored identically to the shape representations, using the *OrdImage* and *ORDImageSignature*.

New images can be added using the *ImportImage* function of the *DBManager* class. This function is almost identical to the *NewRepresentation* function, shown in Figure 31, above.

The actual mapping between thesaurus objects and the image collection was identified as one of the most important, and difficult, aspects of the shape thesaurus framework. Chapter 3.4.3 presented some possible approaches to this. How to create and maintain these links was not one of the principles we wanted to evaluate with the thesaurus prototype, and henceforth no time was spent on developing automatic link generation.

Links between images and thesaurus objects were created manually. The consequences of this are discussed in chapter 6.2.2

Approximately half the images in the image collection were analyzed and linked to the thesaurus objects. The links are represented as the *"Exists_In/Contains"* relationship in Figure 23. Figure 32 presents the structure and a view of the actual data from this table. Links were inserted using the "GenerateLink" method of the "ThHandler" class.

T_OB_NR	BILDE_NR
3	29
3	45
3	50
3	178
4	17

Figure 32 - The "Tbl_Exists_In" describing links between the thesaurus and the images.

The actual image collection is discussed in chapter 5.2.1 and the images can be found in appendix E.

4.2.3 Shape Similarity Function

The built-in similarity functions in O9i were used as a basis for all syntactic feature based searches. A basic presentation of how similarity comparison is performed in O9i is presented here, as it will be useful in understanding the different functions described in the following chapters.

The actual algorithms used in O9i similarity comparison are not public knowledge, as these also are considered business secrets. A brief presentation of the techniques is presented by Guros (2004):

> *Matching of images in interMedia is based on extracting color, texture and shape features from the image and comparing them. The features are extracted by first segmenting the image into regions according to color, then determining the features for each region. Color and texture information are also determined globally (by unifying the region based information) to generate global color and texture histograms. Both the region and global information are stored in the signature.*
>
> *The underlying distance measure for comparison is the Manhattan distance, which is enhanced and combined with a variety of other measures depending on the feature.*

Figure 33 contains the code for the *QBESearch* of the *SearchEngine*class. It is presented here as an illustration of how O9i performs similarity searches. The important parts are highlighted with bold text.

The actual similarity function used is the *IMGSimilar*[12] from the *OrdSys* class. It is based on comparing the signature of a seed image to the signatures of other images. A score is calculated for each image based on how different it is to the seed image. A score of "0" represents identical images.

```
member function QBESearch(seedimage number) return imagelist is
-----------------------------------------------------------------------------------------
-- This function will take the number of an image in the table tbl_seedimage, compare it to
-- the image collection, and return  a list of images with a certain degree of similarity
-----------------------------------------------------------------------------------------

weights  varchar2(64) := 'color="0,33" texture="0,33" shape="0,33" location="0,5"';
treshhold number := 20;

imagenumber number;
score number;
ilist imagelist := imagelist();
counter integer := 1;
compare_sig ordsys.ordimagesignature;

cursor getimages is
  select inumber, ORDSYS.IMGScore(123) SCORE from tbl_images b
    WHERE ORDSYS.IMGSimilar(b.signature, compare_sig, weights, treshhold, 123)=1
    order by score;

BEGIN
select p.image_sig into compare_sig from tbl_seedimage p
  where p.inumber = seedimage;

open getimages;
loop
  fetch getimages into imagenumber, score;
  exit when getimages%NOTFOUND;
  ilist.extend;
  ilist(counter):= imagenumber;
  counter := counter +1;
end loop;

return(ilist);
end QBESearch;
```

Figure 33 - The *QBESearch* method. An illustration of O9i CBIR Search.

The function has two important parameters in addition to the seed signature. The first is a set of weights describing both the relative importance of the syntactic features (Colour, texture and shape) to each other as a percentage as well as how much weight should be given to the spatial location of the features. The first three have to add up to 100%, while the last can be given any weight from 0 to 100%. In Figure 33, these weights are represented in the "weights" string. In this example, each feature is given equal weight, 33%. Location has been set to 50%[13].

Next, a *threshold* determines how similar the signatures must be in order to be included in the result set. Images with a similarity score higher than the threshold

[12] In addition to the IMGSimlar function, ORDSYS class contains another function used to compare images, the IMGScore function. Both are based on the same underlying algorithms, and differ only in the way they can be used to rank and present comparison results. Both are used by the VORTEX Prototype.

[13] The actual usefulness of the "Location" weight has yet to be determined. Experimentation with different values for this has not resulted in any significant difference in search results.

value will be omitted from the search results. Determining a suitable threshold value is difficult, as the similarity scores returned by the similarity comparison varies with the both the weights used and the actual images involved. At times, the actual distance in score *between* a set of images may give a clearer indication of search results than their actual scores. Because of this, using the threshold value to limit search results is not always the best solution.

Finally, a *cursor* is used to store the results of the image search. A *cursor* is a result set in the form of an array, allowing for moving back and forth through the results of the query. In Figure 33, the cursor contains the image number and the calculated score, ordered by score, having images with the lowest score first.

4.3 VORTEX Image Retrieval

VORTEX supports image retrieval based on both "standard" syntactic feature based image retrieval, as well as image retrieval aided with a shape thesaurus. The standard image retrieval, hereafter named OCBIR (OCBIR) was illustrated in Figure 33, above, and is not given any further presentation here. This chapter is focused at describing how VORTEX uses the shape thesaurus for image retrieval.

For the purposes of this experiment, seed images were imported and stored in a separate table, *TblSeedImages*. Each seed image were imported and analysed, and an *OrdImageSignature* feature descriptor was generated. These signatures were used as a basis for OCBIR and thesaurus image retrieval searches.

4.3.1 The Thesaurus Search at a Glance

Figure 34 and Figure 35 shows the *UML Collaboration diagram* and *UML Sequence diagram* for the use case *Search – Thesaurus enhanced QBE*. Figure 34 gives a conceptual overview of the search algorithm and the collaboration between the different objects, while Figure 35 illustrates the actual classes and methods used by the algorithm. The actual search process follows the sequence described in these figures. The search assumes that the query image already exists in the *Tbl_Seed_image* table, and that the primary key of the seed image is used as query input. This is used to retrieve the *OrdImageSignature* of the seed image, which is used to start the search process.

The search process consists of Query Interpretation, performed by the *FindTerm* function, Query Expansion, performed by the *FindRelated* and *FindRelatees* functions, and image retrieval, performed by the *FindLinked*, *FindSimilar*, *QBESearch* and *IMGScore* functions. Result ranking is performed continuously by the *QBTSearch* method. The actual code for the *QBTSearch* function can be found in Appendix C. The functions and procedures used by QBTSearch are detailed in the following subchapters. They can also be found in Appendix C.

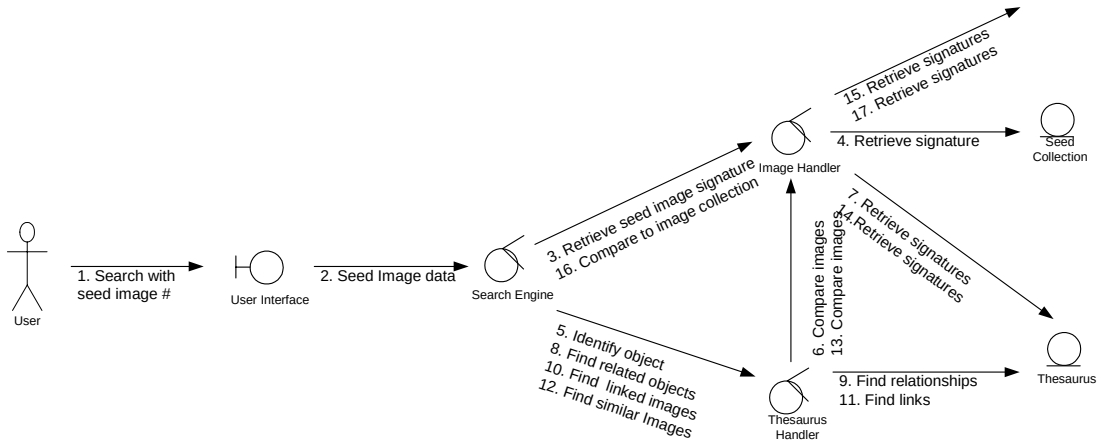


Figure 34 - UML collaboration diagram illustrating the thesaurus-aided image retrieval process.

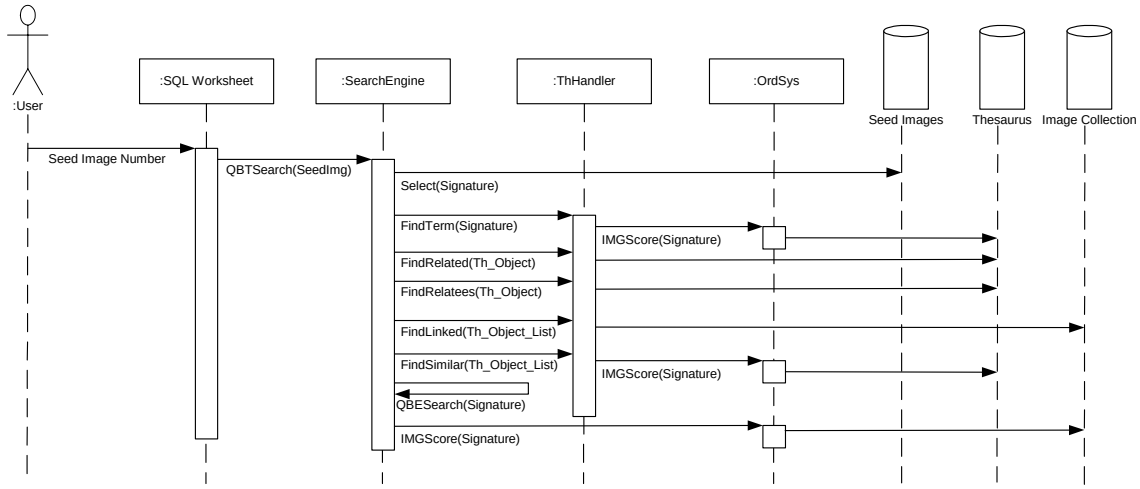


Figure 35 - UML Sequence Diagram showing the search process with the shape thesaurus.

A graphical user interface was not created for the VORTEX prototype, since this would not be required for testing the core functionality. Therefore, interaction with the system is command-line based. Figure 36 shows how a search is initiated. First, an instance of the *SearchEngine class* is instantiated. The actual search is performed by referring to the *QBTSearch* function of the *SearchEngine* class, with "1" as a parameter (referring to the first seed image), and the result of the search is stored as an *ImageList* in the *results* variable. Finally, the *results* list is dumped to the screen.

```
declare
se searchengine := searchengine(null, null);
results imagelist := imagelist();
begin

results := se.QBTSearch(1);
dbms_output.put_line('Image Number');
dbms_output.put_line('============');

for j in results.first..results.last LOOP
    dbms_output.put_line(results(j));
END LOOP;
END;

Thesaurus search completed. Found 83 images:

Image Number
============
6
85
87
[truncated]
```

Figure 36 - Executing a shape thesaurus image search with VORTEX.

4.3.2 Query Interpretation

According to the definition of query interpretation in with a shape thesaurus in chapter 3.4, it should support *visual-to-visual*, *text-to-visual* and *visual-to-text* queries. However, the requirement specification only describes the former. This was considered to be the core functionality of query interpretation, and if it could be shown that this could work as intended, the other query types would be easy to implement.

The actual object identification is handled by the *FindTerm* method of the *ThHandler* class, illustrated in Figure 37. Similarity is based on the *OrdImageSignature* of the seed image used as input to the method. This is compared to the thesaurus shape representations using the *IMGSimilar* method of the *ORDSYS* class, marked by the second line of bold text.

The similarity function is based on the weights specified in the *weights* text string, marked by the first bold line. This describes which syntactic features should be given most weight when calculating similarity score. During the implementation, different weight combinations were tried. The most obvious choice, using the *shape* feature alone, did not produce very meaningful results. It would appear that the *IMGSimilar*[14] function requires some colour and texture description in order to be able to perform shape matching. Different combinations were tried, and the combination of 10% colour, 10% texture and 80% shape was found to give best results.

[14] For more information about the *IMGSimilar* function, see Ward, R. (2001) Oracle interMedia User's Guide and Reference. Accessed: August 1st, 2004. Available Online: http://download-west.oracle.com/docs/cd/B10501_01/appdev.920/a88786/title.htm

```
member function findterm(signature ordsys.ordimagesignature) return number
is
--------------------------------------------------------------------------------
-- This procedure attempts to find the thesaurus representation with the highest
-- similarity to a seed image. Will rank the images, find the one with the lowest
-- score and return this. Hopefully, there will be a match between the seed image
-- and the thesaurus object identified
--------------------------------------------------------------------------------

weights varchar2(64) := 'color="0,1", texture = "0,1", shape = "0,8", location = "1,0"';
t_objnr number;
t_repnr number;
threshold number;
score number;
objects thoblist;
objname varchar2(64);

-- Generate a cursor for use when searching through the collection.
CURSOR getphotos IS
   SELECT t.obj_nr, t.repnr, ordsys.imgscore(123) SCORE
   FROM tbl_thesaurus_rep t
   WHERE ORDSYS.IMGSimilar(t.shape_SIG, signature, weights, threshold, 123)=1 order by
score;

BEGIN
    -- Set Threshold value
    threshold := 100;

   -- Find images matching the profile
   OPEN getphotos;
   LOOP
      FETCH getphotos INTO t_objnr, t_repnr, score;
      EXIT WHEN getphotos%NOTFOUND;
      EXIT;
   END LOOP;
   CLOSE getphotos;
   return t_objnr;
end findterm;
```

Figure 37 - The *FindTerm* method, used to identify thesaurus objects. Important functionality is highlighted.

Finally, the *threshold value* of the similarity function is set to 100. The *IMGSimilar* function had a very high degree of variation in its perceived similarity score, and it was difficult to settle on a *good* threshold value. In most cases, a threshold value at around 50 were sufficient, but in some cases the function managed to correctly identify thesaurus terms, but with a very low similarity score. By setting the threshold value to 100, the similarity function will almost *always* return a possible thesaurus object, even if it is perceived as very dissimilar to the seed image.

The result of query interpretation is a *ThobList* containing only the identified thesaurus object. The list of thesaurus objects is expanded during the Query Expansion phase, and is used as a basis for image retrieval.

4.3.3 Query Expansion

Query expansion is based on finding the thesaurus objects which are *related* to the thesaurus object identified by the query interpretation process. In the current VORTEX prototype, this query expansion is limited to retrieving *all* thesaurus objects

one step removed from the identified object. There is no ordering of the different relationships. As identified in chapter 3.4.2, query expansion could, and most likely ought to, be based on input and feedback from the user; both to as how many steps the query should be expanded, and with respects to how the relationships should be ordered. However, user interaction and feedback has not been implemented in the VORTEX prototype, and was not the main focus of this project. Therefore, this approach was deemed sufficient for the purposes of testing the framework.

Actual query expansion is achieved through the *FindRelated* and *FindRelatees* methods of the *ThHandler* object. The two functions are practically identical, differing only in that the former, illustrated in Figure 38, finds objects that *are related to* a specified thesaurus object, while the latter retrieves objects which specified object *relates to*. Both methods return a list of thesaurus objects, representing the objects which are related to the specified term. In its current implementation, the shape thesaurus functionality in VORTEX does not discriminate between the different relationship types.

```
member function findrelated (thObj number) return thoblist is
--------------------------------------------------------------------
-- This function will find which thesaurus objects are related to
-- the object given as input to the function. It returns a list of
-- numbers, representing the primary keys of the thesaurus objects
--------------------------------------------------------------------

objekter thoblist := thoblist();
objnr number;
counter number := 1;

cursor objects is
  select t_ob_b
  from tbltob_related
  where t_ob_a = thObj;

begin
for objnr in objects loop
  objekter.extend;
  objekter(counter) := objnr.t_ob_b;
  counter := counter + 1;
end loop;

return objekter;
end findrelated;
```

Figure 38 - The *FindRelated* function, used to identify related thesaurus objects.

The thesaurus objects retrieved by the *FindRelated* and *FindRelatees* are added to the *ThobList* created by the *FindTerm* method. The result is a list of thesaurus objects, with the identified thesaurus object first followed by the thesaurus objects related to the identified object.

4.3.4 Image Retrieval and Result Ranking

The actual retrieval of images is based on the *ThobList* containing the thesaurus objects identified during query interpretation and expansion. The retrieval is achieved with three different methods; retrieving images *linked* to a set of thesaurus objects, similarity comparison between a set of object shape descriptors and a similarity comparison between the original query image and the image collection.

Image ranking are performed simultaneously to image retrieval, as the images are ranked in the order they are retrieved:

1. Images known to contain the identified thesaurus object
2. Images known to contain objects related to the thesaurus object
3. Images containing shapes that are similar to the descriptors of the identified object
4. Images containing contain shapes that are similar to the descriptors of the objects related to the thesaurus object.
5. A small set of images which have been found to be structurally similar to the seed image used.

The first four steps are performed based on the objects contained in the *ThobList*. The list is processed two times. During the first processing, images *known* to contain the defined objects are retrieved by retrieving images that are *linked* to the thesaurus objects. This is performed by the *FindLinked* method in the *ThHandler* class, shown in Figure 39 below. This consists of simple SELECT statements performed on the *Tbl_Exists_In* table. The primary key of all images linked to the thesaurus object are retrieved and added to an *ImageList*, called *EndList*.

```
member function FindLinked(thobj number) return imagelist is
--------------------------------------------------------------------
-- This function will find which images are linked to a thesaurus
-- term sent to the function. It will return a list of numbers,
-- representing the primary keys of the images in the image table
--------------------------------------------------------------------

  images imagelist := imagelist(100);
  counter number := 1;
  imagenr number;

  cursor image_search is
    select bilde_nr
    from tbl_eksisterer_i
    where t_ob_nr = thobj;
begin


for objnr in image_search loop
  images.extend;
  images(counter) := objnr.bilde_nr;
  counter := counter + 1;
end loop;

return images;
end FindLinked;
```

Figure 39 - The FindLinked function of the ThHandler object.

During the second processing, images which contain shapes that are *similar* to the shape descriptors representing the thesaurus objects are retrieved. This is performed by the *FindSimilar* method of the *ThHandler* class, shown in Figure 40, below. For each thesaurus object in the list, every shape representation is compared to the image collection, and for each of the shape representations, the 5 images with the highest similarity score is retrieved and added to the *EndList*. The actual similarity comparison is again performed by the *IMGSimilar* function in the *OrdImageSignature* class.

```
member function FindSimilar (termlist thoblist) return imagelist is
-------------------------------------------------------------------------
-- This procedure will take a set of thesaurus terms, find images which
-- are similar to these terms, rank them and return a list of images
-- which is perceived to be similar to the thesaurus terms in shape.
-------------------------------------------------------------------------

-- Variables
images imagelist := imagelist();
compare_sig ordsys.ordimagesignature;
counter number := 1;
treshhold_number := 25;
limit_number := 4;
tterm number;
imagenumber number;
teller number := 0;
dummy_score number;
weights varchar2(64) := 'color="0,1", texture = "0,1", shape = "0,8", location = "1,0"';

-- Cursors
cursor getimages is
  select inumber, ordsys.IMGScore(123) SCORE from tbl_images b
  WHERE ORDSYS.IMGSimilar(b.signature, compare_sig, weights, treshhold, 123)=1 order by score;

cursor getrepresentations is
select shape_sig from tbl_thesaurus_rep
where obj_nr = tterm;

-- Main body
BEGIN

-- Loop through the termlist
for n IN 1..termlist.count LOOP
  tterm := termlist(n);
  dbms_output.put_line('Working on term ' || tterm);

  --- Loop through the representations of the thesaurus object
  open getrepresentations;
  loop
    fetch getrepresentations into compare_sig;
    exit when getrepresentations%NOTFOUND;

    --- Loop through the images similar to the representation image
    open getimages;
    loop
      fetch getimages into imagenumber, dummy_score;
      exit when getimages%NOTFOUND;
      images.extend;
      images(counter) := imagenumber;
      counter := counter +1;
        teller := teller + 1;
        if teller > limit_number then EXIT; end if;
    end loop;
    close getimages;
    --- Finished with the first representation of the object
  end loop;
  close getrepresentations;
  --- Finished with all representations of this object
END LOOP;

--- Finished with the entire loop.
-- Return
return images;
END FindSimilar;
```

Figure 40 - The *FindSimilar* function in the *ThHandler* class.

Limiting the above searches to a 5 images was based on experimentation with threshold values and different limits. Initial testing with no limitations resulted in very large result sets, returning almost all the images in the image database in one extreme case. Next, different threshold values were tried. The calculated similarity score was

very dependent on the structure of both the seed images and the images in the collection. In some cases, using a high threshold value would retrieve almost all images in the collection, while returning no images in another case. Furthermore, a low threshold value would exclude images that were clearly both similar and relevant. The best overall results were achieved using a high threshold value and focusing more on the actual order of the retrieved images.

Finally, step 5 is performed by using a standard OCBIR search, using the *QBESearch* method belonging to the *SearchEngine* class, shown in Figure 33, page 62. Again, similarity search is based on using the *IMGSimilar* methods. However, here the weights are evenly divided between colour, texture and shape, as the goal is to find images that are similar to the seed image, and not necessarily only the identified thesaurus objects. Any images found here are added to the end of the *EndList*.

The result of the image retrieval process is a ranked list of images, represented by their primary key, which can be used for actual presentation of the search results. In the current VORTEX prototype, it is dumped to the Oracle user interface as a list of primary keys.

5 Evaluating the VORTEX Prototype

The *Shape Thesaurus* framework described in chapter 3 was partly implemented in the *VORTEX* prototype described in chapter, and has been shown to be able to retrieve images with the aid of a Shape Thesaurus. With that, we have shown that it is possible to implement the suggested framework.

In order to evaluate the retrieval properties of the prototype, it has been tested through an experiment in which it has been compared to a simple CBIR system, represented by the built in CBIR retrieval functions provided by Oracle 9i interMedia.

5.1 Experimental Design

5.1.1 Experiment Classification

The research project described in this thesis can be classified as a laboratory experiment, or as having an *experimental design*. A classic laboratory experiment consists of examining two groups of units, where one is the *experiment unit* and the other the *control unit*. When we do an experiment using scientific method, we are interested in the effect that a method or tool, called a *factor*, has on an *attribute of interest*. Each agent that we study and collect data on is called a *subject* or an *experimental unit*. The goal of an experiment is to collect enough data from a sufficient number of subjects, all adhering to the same treatment, in order to obtain a statistically significant result on the attribute of concern, compared to some other treatment. Table 6 presents an illustration of how the project described in this thesis can be classified as an experiment:

Table 6 - Experiment Classification.

Experiment component	This project
Experiment Unit	Image Retrieval with a Shape Thesaurus
Control Unit	Image Retrieval based on standard OCBIR
Factor	A Shape Thesaurus
Attribute of Interest	Image Retrieval Efficiency in terms of recall and precision
Experiment Unit	Query set

In this thesis, image retrieval with a shape thesaurus is compared to image retrieval based on syntactic features alone, represented by the similarity functions provided by Oracle 9i interMedia. Both image retrieval approaches are used on a simple image retrieval system. The first approach might be considered equivalent to the *experiment unit*, while the latter a *control unit*. The shape-thesaurus framework is the *factor* of analysis, and the retrieval result is the *attribute of interest*. Data on the retrieval results is collected through a group of queries, which can be considered as the *experiment subjects*, and the collected data is used to perform an evaluation of the attribute of concern.

5.1.2 Experiment Goal and Design

The main purpose of this experiment was to answer the research question and hypothesis for this project, repeated here:

Research question

> *Can recall / precision measures for an image retrieval system be significantly improved by utilizing a thesaurus for shapes?*

Hypothesis

> *An image retrieval system that utilizes a thesaurus for shapes, will lead to a significant improvement in recall / precision results over a system based on syntactical feature comparison.*

The VORTEX system contains one possible implementation of a Shape Thesaurus, and has been used in the experiment as an instrument for evaluating the research question through the hypothesis. We have already seen that VORTEX is capable of utilizing a shape thesaurus for image retrieval, thus the experiment is aimed at comparing its capabilities to an "ordinary" feature-based image retrieval system. A system for comparison was developed using the built-in functionality in Oracle 9i interMedia, hereafter named OCBIR.

In chapter 2.2 we presented five different levels retrieval which should be supported by an image retrieval system. However, in chapter 3.1 we identified that the main focus of the shape thesaurus is to improve identification and retrieval of *visual objects*. As a result, it would not be very meaningful to examine queries of level 0 (external features) and level 1 (Syntactical features). Furthermore, level 2 queries (Generic features) were identified as the main focus of the shape thesaurus. Finally, as the shape thesaurus is focused at identifying *objects*, it is not very well suited for identifying abstract content other that activities and scenes based around identifiable objects. Based on this, it was decided to classify image retrieval tasks into the following categories:

- Retrieval based on *generic objects*
- Retrieval based on *generic scenes*
- Retrieval based on *specific objects and scenes*
- Retrieval based on *abstract objects, scenes and activities.*

The two first categories represent generic features, which are the main focus for the shape thesaurus. The two last categories were included in order to gain an understanding as to *how far* the shape thesaurus could be able to assist during image retrieval activities. The levels represent increasingly difficult information requests, and were used as a basis for classifying the queries used as experimental units.

Furthermore, in chapter 2.4.3 visual image retrieval were classified as based on *Query-By-Sketch* and *Query-By-Example*. Both query types can be used with both standard, syntactic based image retrieval techniques, as well as with the shape thesaurus. Therefore, both query types were evaluated in this thesis. This results in a total of eight different categories for comparison, as shown in Table 7, below. In addition to these categories, overall results for all levels for QBE and QBS, as well as overall results for both systems, were evaluated.

Table 7 – Experimental evaluation categories.

Retrieval Level	Query type	
	QBE	QBS
Generic Objects	**A**	**B**
Generic Scenes	**C**	**D**
Specific Objects and Scenes	**E**	**F**
Abstract objects, scenes and activities.	**G**	**H**

The actual experiment consisted of the following steps:

1. Build a collection of images used as a test collection
2. Generate a set of verbal queries for use as a basis for retrieval measurement
3. Use an external test group to express the queries visually through QBE and QBS
4. Use the external test group to identify images relevant to each query
5. Execute the queries using OCBIR and the VORTEX Shape thesaurus
6. Measure the query results using recall and precision
7. Evaluate recall / precision as a measure of image retrieval efficiency through significance testing.

In addition to this, a questionnaire was submitted to the test group in order to gather data that might provide additional information. As this was not central to the research question, less focus has been given to this than measurement and evaluation of recall and precision. The questionnaire was introduced as it proved an interesting opportunity to retrieve information that could not be extracted from the recall / precision measurements; how *users* felt about expressing queries visually, through example images and custom drawings.

5.2 Experimental Framework

5.2.1 Test collection – Image Database

In order to perform the experiment, a set of images was collected from various sources on the internet[15]. A total of 196 images were included. The collection was made in accordance with the domain for the shape thesaurus; a collection of images depicting marine animals and related activities.

The number of images has been kept small in order to maintain a clear overview of which images are relevant results for each query. Images were selected based on their suitability for the experiment; some images are very similar in both structure and semantics, others are similar in structure but different in semantics, and vice versa.

Furthermore, the images have varying degrees of semantic complexity, as illustrated in Figure 41. The first image is a very basic image; a stylistic drawing of a dolphin, clearly silhouetted against the background. The second image is still basic, one

[15] About two thirds of the images were gathered from National Oceanic and Atmospheric Administration (http://www.photolib.noaa.gov/) and The National Maritime Museum (http://www.nmm.ac.uk/), while the rest were gathered from various internet sites.

defined shape against a relatively stable, homogeneous background. The last two images are more complex. The first depict a killer whale, possibly named, involved in an activity, against a heterogeneous background. The last image contains two objects which might be difficult to tell apart.

The choice to use a custom rather than existing image collection was based on availability; it was difficult to get access to a suitable image collection. The entire collection is available in Appendix E.

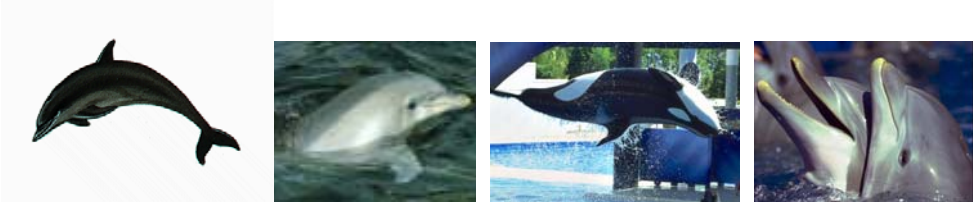


Figure 41 - Image examples from the test collection.

5.2.2 Query Set and Specification

The query set was developed based on the four levels identified in Table 7. For each of these levels, a set of queries were created, representing different information needs in the same semantic level. The actual queries are illustrated in Table 8, below.

The query set was developed after the image collection was created, and reflects the actual collection. They were created as a representation of some of the possible information requests users might present to this image collection. The actual visual expression of the queries as query images was performed by an external test group.

The first 6 queries are focused at retrieval of images containing generic objects, such as whales, dolphins and (human) divers. These queries are representative of information needs where the enquirer wishes to retrieve generic images of a certain object. The next 7 queries are focused at retrieval of images depicting generic scenes, such as a pair of dolphins, or birds against the backdrop of the sky. Together these two sets of queries represent the main focus of the shape thesaurus; queries aimed at *generic* semantic image content.

Queries 14 – 16 are focused on retrieving images containing a specific objects or specific scenes. Ordinarily, this would represent queries after named (unique) objects or scenes, such as “the New York Skyline”, “Bergen Wharf” and “the killer whale Keiko”. However, neither the structure of the shape thesaurus, nor the image collection itself lends itself well to queries of this type. First of all, the thesaurus and its shape descriptors are focused on generic objects rather than specific individuals. Furthermore, there are not many easily recognizable marine animals other than the killer whale “Keiko”. As a result, there only three queries have been included in this category. Query 14 is an attempt to find images containing the killer whale “Keiko”, representing a specific object. Queries 15 and 16 represent queries against specific scenes. The information need behind the queries is presented as an image, and the queries are attempts to find the scenes depicted in these images. The images are shown in Figure 42, below.

The final query set (Queries 14 through 24) represent abstract information need. As the focus of the shape thesaurus is to recognize objects rather than abstract content and meaning, the queries are focused on retrieving images containing visual objects (animals, people or items) involved in some sort of activity, rather than on intrinsic content and derived meanings.

Table 8 - Query set developed for the experiment.

Query #	Query
	Generic objects
1	Find images depicting a 'whale'
2	Find images depicting a 'dolphin'
3	Find images depicting a 'shark'
4	Find images depicting a 'bird'
5	Find images depicting a 'seagull'
6	Find images depicting a 'diver'
	Generic scenes
7	Find images depicting 2 or more 'dolphins'
8	Find images depicting 2 or more 'seagulls'
9	Find images depicting 'animals' on the surface
10	Find images depicting one or more 'whales' on the surface
11	Find images depicting one or more 'birds' in the sky
12	Find images depicting 'animals' or 'divers' under water
13	Find images depicting both a 'diver' and an 'animal'
	Specific objects and scenes
14	Find images depicting the killer whale 'Keiko'
15	Find the images of the scene depicted in image 106
16	Find the images of the scene depicted in image 160
	Abstract objects, scenes and activities
17	Find images depicting one or more 'dolphins' 'playing'
18	Find images depicting one or more 'dolphins' 'jumping'
19	Find images depicting one ore more 'whales' 'jumping'
20	Find images depicting one or more 'seagulls' 'eating'
21	Find images depicting one or more 'sharks' 'attacking'
22	Find images depicting a 'dolphin' 'playing' with a 'ball'
23	Find images of the killer whale 'Keiko' being 'fed' by a 'man'
24	Find images of 'humans' and 'animals' 'interacting'

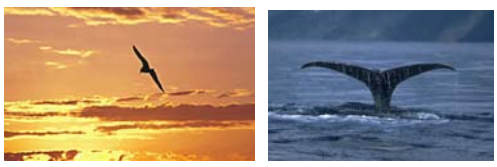


Figure 42 - Images describing query 15 (left) and query 16 (right).

5.2.3 Test group

One of the implications of using a custom-built image collection is that I had a complete overview of the images in the collection. This could potentially lead to the seed images being biased. Likewise, determining which images are relevant to a search might be biased by knowledge of the capabilities of both the two retrieval algorithms implemented in VORTEX. In order to reduce this bias, a group of external persons was included in the experiment. The people in the test group had no prior knowledge of the image collection or the capabilities of the two retrieval algorithms to be evaluated, which removed the possibility of biased queries and predetermination of relevant images.

The final group consisted of six people, 3 men and 3 women, who volunteered to participate in the experiment. Three were fellow students with some basic knowledge of image retrieval systems. The other three were external persons, with little or no knowledge of the challenges described in this thesis. All were between 20 and 32 old, and had different levels of computer proficiency.

5.2.4 Questionnaire

In addition to providing the experiment with query images and a list of relevant images for each query, the test group given a questionnaire regarding how they felt about using example images and drawings to express information requests. Although not directly linked to the research question or hypothesis, it proved to be an excellent opportunity to collect data which otherwise might be lost.

The questions asked in the questionnaire were divided into three categories. The first consisted of a set of questions answerable on a 6-point Likert scale:

1. How easy / difficult was it to find a seed image for each of the queries
2. How easy / difficult was it to express the queries using a drawing
3. How close do you feel your example image matched *the most relevant* image in the image collection?
4. How close do you feel your drawn image matched *the most relevant* image in the image collection?
5. Overall – how easy was it to find good example images?
6. Overall – how easy was it to draw good example images?

The first two questions were related directly to the different queries, while two last questions were included in order to let the respondents give their own opinion of their *overall* feeling towards the two query specification methods. Question 3 and 4 were included in order to examine if there were any discrepancies between what the respondents *felt* should be retrieved, and the actual images retrieved.

The next set of questions included questions answerable in plain text:

7. If you had trouble finding good seed images for any of the queries: Please note which queries were difficult, and, using your own words, describe what the difficulties were.

8. If you had trouble creating good drawings for any of the queries: Please note which queries were difficult, and, using your own words, describe what the difficulties were.

The last two questions were allowed the respondent to give some general information about using Query-by-Example

9. Have you tried searching for images using Query by Example before?

10. Any other comments to using Query-By- Example when searching for images?

The actual questionnaire is available in Appendix F.

The main purpose of the questionnaire was to attempt to gain an understanding about how users feel about using images to express an information need. Is there any difference between general, specific and abstract queries? And how do the actual query results correspond to how well the respondents perceived their own queries to be?

As it stands, the questionnaire is not very well put together, and it might be difficult to use it to make any conclusions about the issues raised. However, evaluation of the answers hinted at some interesting findings, which are briefly discussed in chapter 5.3.5

5.3 Experiment Execution – Data collection

5.3.1 Query Specification and Test Group

The first part of the experiment was the actual data collection. Initially, the queries described in Table 8 were doubled and divided into eight subsets consisting of 6 queries each, representing 8 respondents. However, only 6 respondents were available and not all queries could be doubled. Since there were fewer queries in the third group, it was decided that all these should be doubled. This resulted in an unequal distribution of queries between the respondents; respondent 5 did not get any queries representing abstract content. This has the potential to introduce a bias at least in the result set for the different query levels, if not for the overall results.

Figure 43, below, presents an overview of the queries were divided between the respondents. In the following, a query which has been doubled has been labelled with an “A” for the first query and a “B” for the second; i.e. “1A” and “1B”.

The respondents were given two tasks:

- Express the set of queries through an *example image* and through a *sketch* drawn by themselves – representing Query-By-Example and Query-By-Sketch
- Evaluate which of the images in the collection *they* found to be relevant to the queries they were given.

For the example images, they were given total freedom to use any digital image from any source. An oral survey after the experiment revealed that all respondents used *Google* to find images.

For the image drawings, they were asked to format the images as black and white drawings on a 400 by 400 pixel image. Most of the example images were kept to this format, with some exceptions. This was done to have the drawings as close to the shape templates as possible, simulating a user interface adapted to this. All the respondents used Paint Shop Pro 8 to create the drawings, except one who used Microsoft Paint. The example images and drawings the respondents supplied can be found in Appendix G.

Query #	Respondent					
	1	2	3	4	5	6
1	x				x	
2		x				x
3			x			
4				x		
5				x		
6					x	
7		x		x		
8					x	
9		x				
10	x		x			
11	x					
12						x
13					x	
14	x				x	
15			x			x
16				x	x	
17		x				x
18		x	x			
19	x					
20				x		x
21			x	x		
22			x			
23	x					x
24		x				

Figure 43 - Queries divided into subgroups.

After finishing the queries, the respondents were given a list of all the images in the image collection and were asked to select the images *they* found to be relevant to the relevant to the search. They were asked to rank the images in order of relevance, although this was not done properly by all respondents. This was used as a basis for recall and precision measurement. Finally, they were given the questionnaire. The time spent by the respondents varied from 45 minutes to three hours, with an average of about 90 minutes.

The respondents were given the following handouts during the experiment:

- Information letter describing their task.
- A list of 6 different queries.
- A form for noting which images were relevant to their queries.
- A list of the images in the database.
- A questionnaire.

The image collection is presented in Appendix E. The rest of the handouts are presented in Appendix F.

The respondents were given the handouts both as paper handouts, and electronic copies, delivered on either a CD or over the internet. The set of images was returned by e-mail, and the forms and questionnaires were filled out and returned by hand. The images and forms were stored and classified according to respondent number, and all connections between the collected data and the respondents were removed after checking for any missing data.

5.3.2 Query Execution

The query images submitted by the users were imported to the image collection and stored in a separate database table, Tbl_Seed_image. The actual queries were performed by a simple PL/SQL statement entered through O9i command line interface. The images were first put through OCBIR, using the statements illustrated in Figure 44. Search output was presented as a simple list of image numbers, referring to the primary keys in the image collection table.

```
declare
se searchengine := searchengine(null, null);
results imagelist := imagelist();
begin

results := se.QBESearch(1);
dbms_output.put_line('Image Number');
dbms_output.put_line('============');

for j in results.first..results.last LOOP
    dbms_output.put_line(results(j));
END LOOP;
END;
```

Figure 44 - Executing a query in OCBIR.

Next, the queries were put through the shape thesaurus retrieval algorithm, using a PL/SQL statement similar to the statement in Figure 44, above.

Finally, the both results from both algorithms were organized into an Excel spreadsheet along with the list of relevant images for each query. There were a total of 4 queries executed for each specified query; Query-By-Example and Query-By-Sketch for both OCBIR and the Shape Thesaurus.

Figure 45 illustrates how the results were organized. The data was organized into 4 main segments. The first segment contains information about the actual query; query number, respondent number, seed image numbers for both example image and user drawing, the textual query and the terms identified by shape thesaurus. For this example, we see that the query is “Find images containing two or more ‘Dolphins’”. The shape thesaurus correctly identified ‘dolphin’ from the example image, and identified ‘Dolphin Fin’ from the drawn image.

The next segment represents the relevant images identified by the respondent; number of images and the actual image number. In this example, we see that the respondent has identified 12 images as relevant.

Query Details

Query #	7A	**Query**	Find images depicting 2 or more 'dolphins'
Respondent	2		
		Shape Thesaurus Shape Identification	
Seed Image	3	*Seed Image*	Dolphin
Drawn Image	4	*Drawn Image*	Dolphin Fin

Relevant Images																
Image Number	174	162	65	64	58	56	55	46	44	43	35	32				
Total Relevant																
12																

Retrieved Images Relevant images are marked in **bold** and greyed out

OCBIR																
QBE	18	38	20	119	22	27	62	26	128	140	**64**	139	49	185	137	34
	23	122														

OCBIR																
QBS	203	127	143	63	144	52	187	51	28	201	67	60	101	88	42	193
	66	102	59	12	37	181	103	**174**	77	154	182					

Shape Thesaurus																
QBE	100	9	10	33	**43**	**65**	100	2	31	33	**35**	37	39	41	**46**	45
	47	49	51	53	**55**	57	59	61	63	65	**174**	61	169	151	183	143
	174	203	81	150	203	52	6	169	122	103	81	103	94	119	81	34
	75	150	157	166	75	122	**64**	**44**	83	151	61	103	103	119	15	103
	103	34	203	203	34	18	38	20	119	22						

Shape Thesaurus																
QBS	2	31	33	**35**	37	39	41	**43**	45	47	49	51	53	**55**	57	59
	61	63	**65**	**174**	9	10	33	43	**46**	**64**	128	29	27	170	**44**	83
	151	61	103	103	119	15	103	103	34	203	203	34	61	174	203	81
	150	203	52	6	169	122	103	81	203	127	143	63	144			

Summary

	Retrieved	Relevant	Recall	Precision
OCBIR Seed	18	1	0.08	0.06
OCBIR Drawn	27	1	0.08	0.04
Shape Th. Seed	74	8	0.67	0.11
Shape Th. Drawn	61	8	0.67	0.13

Figure 45 - Example of query result specification, showing query 7A.

Further down is the actual images retrieved by OCBIR and the shape thesaurus. This part is divided in four sub segments; OCBIR QBE, OCBIR QBS, Shape Thesaurus QBE and Shape Thesaurus QBS. The images are listed in the same order as they were retrieved by the retrieval algorithms, from top left to bottom right. The relevant images retrieved are greyed out and marked by bold text.

The last segment is a summary of the four image queries. This shows the number of retrieved images, the number of relevant images retrieved as well as a recall / precision measurement for each of the four query executions. Query details for all 36 query sets are available in Appendix H.

5.3.3 Measurement of Recall and Precision

Recall and precision calculations were made for each individual query, as seen in Figure 45. When evaluating retrieval algorithms on a set of queries, the most used measurement is based on measuring precision on a set of 11 standard recall levels and then generating a recall / precision curve for each query. These curves are then averaged, resulting in an average recall and precision curve for each retrieval algorithm (Baeza-Yates and Ribeiro-Neto 1999; Lu 1999).

However, during the execution of this experiment, it became evident that this approach might not be optimal. Neither OCBIR nor retrieval with the shape thesaurus managed to achieve very high recall levels; the average recall level for OCBIR was 10%, while the average level for retrieval with the Shape Thesaurus was 28%. Although some queries had higher recall, using the 11 standard recall levels would not provide very interesting results. A search through literature did not present any good solutions to this situation. As a result, I decided to customize two standard recall / precision measurements for this experiment.

First, to present the retrieval systems visually, I used *number of relevant images retrieved* as the basic unit of measurement; precision was measured for each of the 10 *first relevant images retrieved*. This was used to create recall / precision curves.

Furthermore, for the actual comparison between the two systems, I used *single value summaries* based on each of the queries used. Both measures are detailed below. For both measurements, average recall and precision were calculated for each of the comparison criterions presented in Table 7 (page 73). While these customized recall and precision measurements makes it difficult to compare the results to other evaluations, they ensure that a thorough comparison between OCBIR and retrieval with a shape thesaurus is possible.

Figure 46 below, showing Query 1B, illustrates how precision for each of the 10 first relevant images retrieved were measured. The upper part of the figure represents the results achieved by CBIR, while the lower part represents the results achieved with the shape thesaurus.

The first column represents the recall level, shown as number of *relevant images* retrieved. The second column shows the total number of retrieved up to that recall level. In the case of OCBIR example image, the first *relevant* image was the fourth image returned. The third column is the actual recall level achieved at this level, calculated on the basis of the actual number of relevant images. Finally, the last

column represents precision at this level. Continuing with the first example, 25% represents that one in four retrieved images are considered relevant.

OCBIR – QBE

Relevant Images	Total Retrieved	Recall	Precision
1	4	1,89 %	25,00 %
2	9	3,77 %	22,22 %
3	11	5,66 %	27,27 %
4	13	7,55 %	30,77 %
5	16	9,43 %	31,25 %
6	17	11,32 %	35,29 %
7	24	13,21 %	29,17 %
8		0,00 %	
9		0,00 %	
10		0,00 %	

OCBIR – QBS

Relevant Images	Total Retrieved	Recall	Precision
1	1	1,89 %	100,00 %
2	3	3,77 %	66,67 %
3	8	5,66 %	37,50 %
4	25	7,55 %	16,00 %
5	28	9,43 %	17,86 %
6	40	11,32 %	15,00 %
7		0,00 %	0,00 %
8		0,00 %	0,00 %
9		0,00 %	0,00 %
10		0,00 %	0,00 %

Retrieval with Shape Thesaurus - QBE

Relevant Images	Total Retrieved	Recall	Precision
1	30	1,89 %	3,33 %
2	31	3,77 %	6,45 %
3	37	5,66 %	8,11 %
4	53	7,55 %	7,55 %
5	59	9,43 %	8,47 %
6	68	11,32 %	8,82 %
7	74	13,21 %	9,46 %
8		0,00 %	0,00 %
9		0,00 %	0,00 %
10		0,00 %	0,00 %

Retrieval with Shape Thesaurus - QBS

Recall Level	Retrieved	Recall	Precision
1	23	1,89 %	4,35 %
2	29	3,77 %	6,90 %
3	35	5,66 %	8,57 %
4	36	7,55 %	11,11 %
5	51	9,43 %	9,80 %
6	53	11,32 %	11,32 %
7	56	13,21 %	12,50 %
8		0,00 %	0,00 %
9		0,00 %	0,00 %
10		0,00 %	0,00 %

Figure 46 - Measurement of recall / precision for query 1B - "Find Images Depicting a Whale".

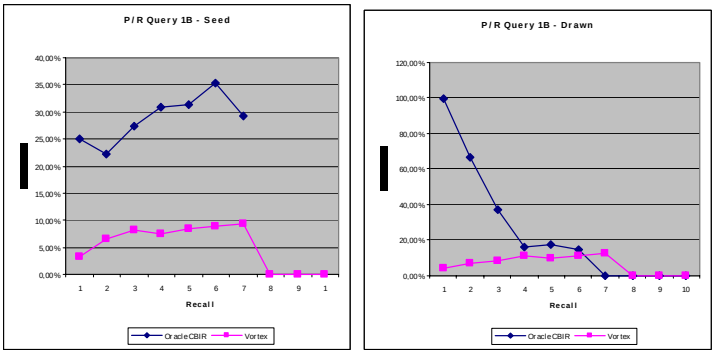


Figure 47 - recall and precision curves for Query 1B. The horizontal axis represents recall as number of images retried, while the vertical axis represents precision.[16]

The precision / recall values for the each query were visualised through precision / recall curves, as presented in Figure 47. The first graph shows precision for the ten

[16] In these figures, "VORTEX" refers to image retrieval with a shape thesaurus.

first images recalled for the example image used in query 1B, while the second graph shows the drawn image. The recall / precision calculations for all queries are available in Appendix J.

The average recall and precision were calculated by summing up the precision at each recall level for all queries, and dividing this by the number of queries. In addition to the above categories, the average of all queries was calculated. Figure 48 shows the average recall and precision values calculated for *all* queries.

OCBIR						
QBE			*QBS*		*Average (QBE + QBS)*	
Recall	Precision		Recall	Precision	Recall	Precision
1	17,81 %		1	16,63 %	1	17,22 %
2	9,04 %		2	6,94 %	2	7,99 %
3	5,10 %		3	5,18 %	3	5,14 %
4	4,96 %		4	1,17 %	4	3,06 %
5	4,58 %		5	1,35 %	5	2,96 %
6	2,56 %		6	1,17 %	6	1,87 %
7	0,81 %		7	0,71 %	7	0,76 %
8	0,00 %		8	0,79 %	8	0,39 %
9	0,00 %		9	0,86 %	9	0,43 %
10	0,00 %		10	0,92 %	10	0,46 %

Shape Thesaurus						
QBE			*QBS*		*Average (QBE + QBS)*	
Recall	Precision		Recall	Precision	Recall	Precision
1	14,29 %		1	14,20 %	1	14,24 %
2	8,72 %		2	12,81 %	2	10,76 %
3	7,94 %		3	10,06 %	3	9,00 %
4	5,19 %		4	9,49 %	4	7,34 %
5	4,15 %		5	7,70 %	5	5,92 %
6	2,94 %		6	6,74 %	6	4,84 %
7	2,25 %		7	6,59 %	7	4,42 %
8	2,06 %		8	5,23 %	8	3,64 %
9	0,00 %		9	4,55 %	9	2,28 %
10	0,00 %		10	3,48 %	10	1,74 %

Figure 48 - Average recall / precision for all queries.

The upper part of the figure is refers to OCBIR. The first segment shows average recall / precision values for all queries based on example images, while the second segment shows averages for queries based on drawn images. The last segment shows the average of both example images and drawings. The averages were visualized through recall / precision curves, as illustrated in the following three figures.

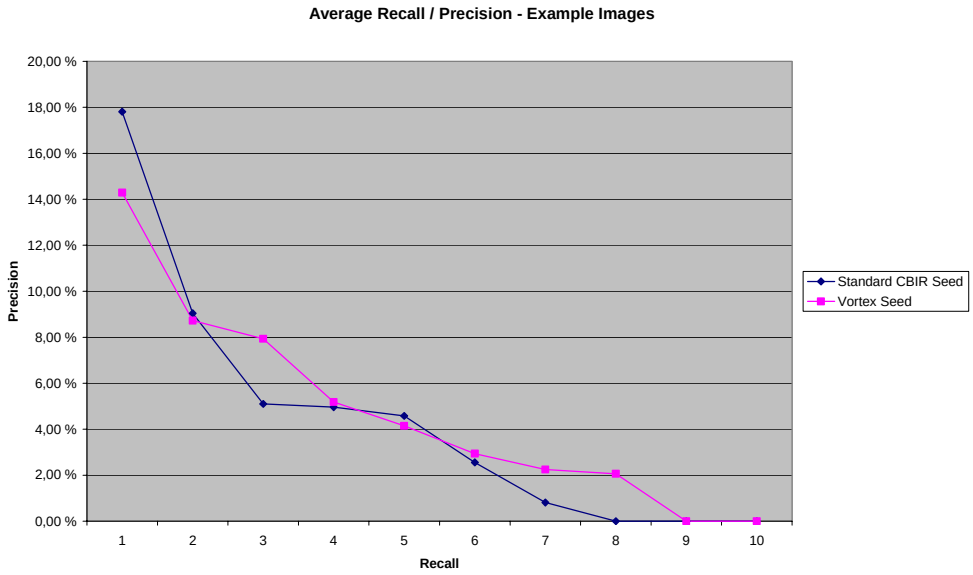


Figure 49 - Average recall / precision for all queries based on example images. The horizontal axis represents recall level shown retrieved relevant images, while the vertical axis represents precision in percent.

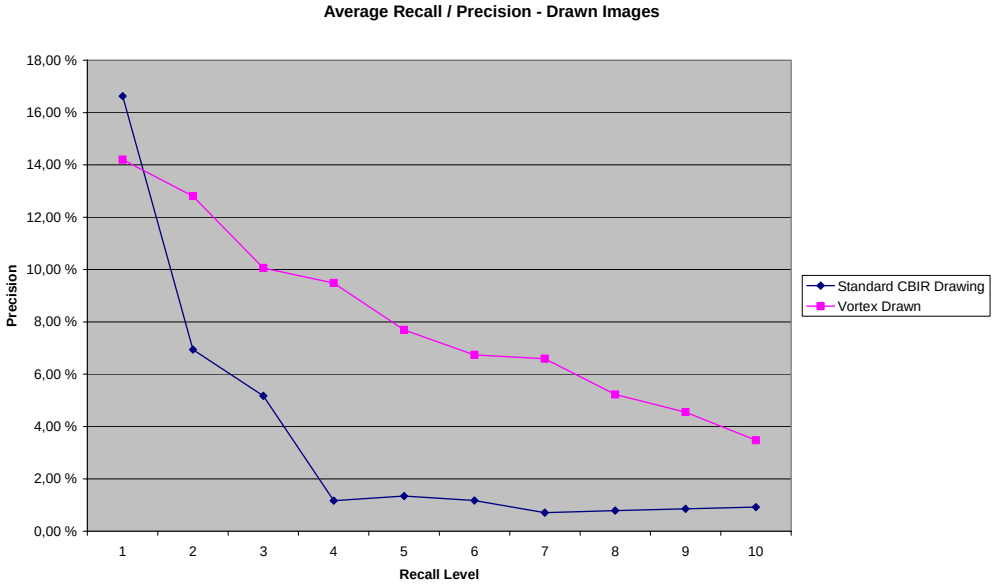


Figure 50 – Average recall / precision for all queries based on drawn images. The horizontal axis represents recall level shown retrieved relevant images, while the vertical axis represents precision in percent.

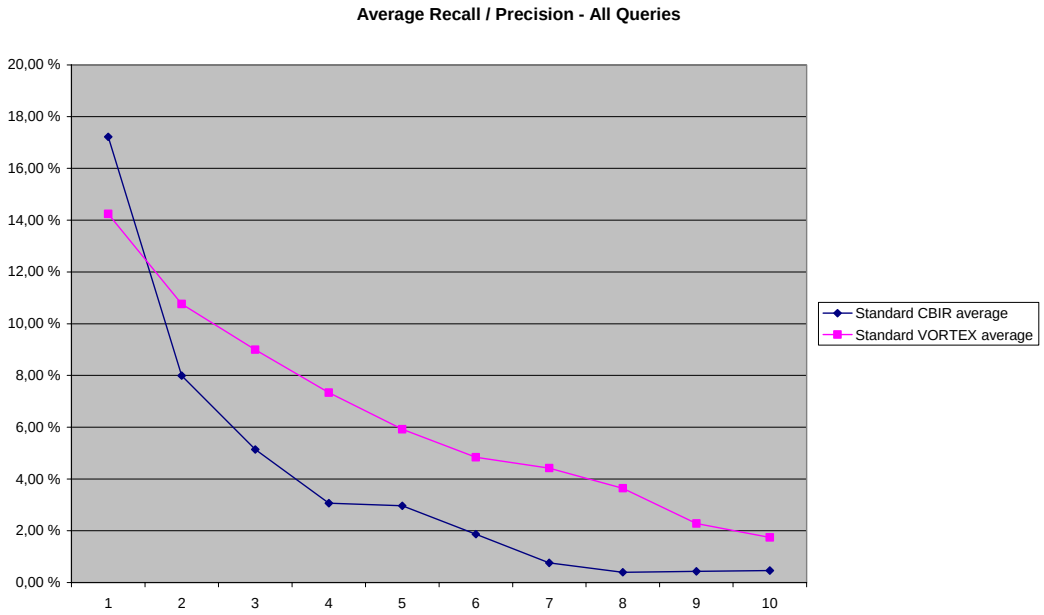


Figure 51 - Average recall / precision for all queries. The horizontal axis represents recall level shown retrieved relevant images, while the vertical axis represents precision in percent.

Finally, Figure 52 shows the average recall and precision curves for both example and drawn images for both retrieval algorithms.

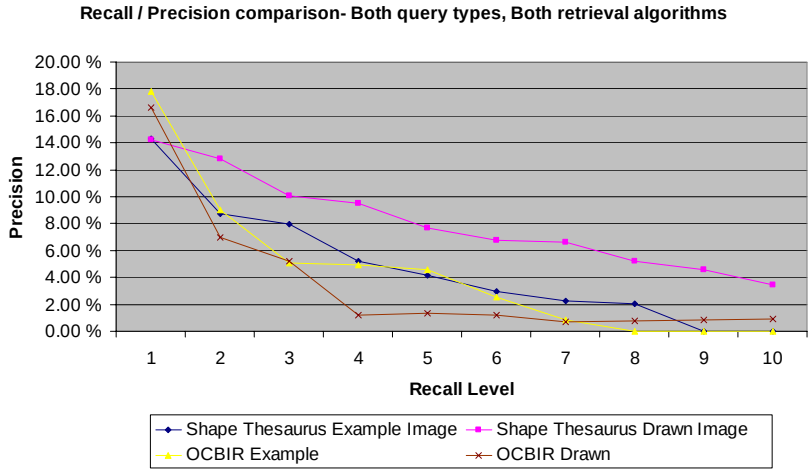


Figure 52 - Average recall / precision for all query types, both retrieval algorithms.

A complete overview of recall / precision averages and their corresponding curves can be found in Appendix J. Actual evaluation of recall / precision values is found in chapter 6.

At first glance, the above curves seem to indicate that retrieval with the shape thesaurus is capable of achieving better retrieval, both in terms of recall and precision, than OCBIR. However, due to the nature of the averaging process, this visualisation

might not give an entirely correct presentation, especially in the case of precision. The scale of the curves is quite small (From 0 to 20% in most cases), and in most cases, precision is quite low. As a result, single queries might have a significant effect on the average curves, especially for the later recall levels, where the precision obtained is very low.

As a result of this, another recall and precision measurement, *single value summaries*, were used. This allows us to examine the actual recall and precision obtained by each query. Two different measures of single value summaries were used; an approach based on *R-Precision* and the *actual recall and precision* values achieved by the two systems for *each query*.

The first approach was used to create *recall* and *precision* histograms as visualizations of the search results from each query. The approach was based on *R-Precision* as presented in (Baeza-Yates and Ribeiro-Neto 1999). However, rather than measuring precision at the standard *R* position, I used the *actual precision* and *actual recall* achieved for each particular query by the two systems. Let

(1) $\mathbf{P_A(}\boldsymbol{i}\mathbf{)}$ and $\mathbf{P_B(}\boldsymbol{i}\mathbf{)}$

and

(2) $\mathbf{R_A(}\boldsymbol{i}\mathbf{)}$ and $\mathbf{R_B(}\boldsymbol{i}\mathbf{)}$

be precision (1) and recall (2) achieved by OCBIR (A) and retrieval with the shape thesaurus (B) for the *i*-th query.

Then define the following values:

(3) $\mathbf{P_{A/B}(}\boldsymbol{i}\mathbf{) = P_A(}\boldsymbol{i}\mathbf{) - P_B(}\boldsymbol{i}\mathbf{)}$

and

(4) $\mathbf{R_{AB}(}\boldsymbol{i}\mathbf{) = R_A(}\boldsymbol{i}\mathbf{) - R_B(}\boldsymbol{i}\mathbf{)}$

as measures for the differences in precision (3) and recall (4) for OCBIR and retrieval with the Shape Thesaurus. A value of $\mathbf{P_{A/B}(}\boldsymbol{i}\mathbf{)}$ or $\mathbf{R_{AB}(}\boldsymbol{i}\mathbf{)}$ equal to 0 indicates that both algorithms are capable of equivalent precision or recall for the *i*-th query. A positive value indicates better performance for OCBIR, while a negative value indicates better performance for the shape thesaurus. Table 11 shows recall, precision, $\mathbf{P_{A/B}(}\boldsymbol{i}\mathbf{)}$ and $\mathbf{R_{AB}(}\boldsymbol{i}\mathbf{)}$ for all queries.

The values $\mathbf{P_{A/B}(}\boldsymbol{i}\mathbf{)}$ or $\mathbf{R_{AB}(}\boldsymbol{i}\mathbf{)}$ were to create *recall and precision histograms*. These diagrams allow us to quickly compare the retrieval performance history of the two systems through visual inspection. Figure 53 and Figure 54 shows the recall and precision histograms for all queries – average of both query types. As with $\mathbf{R_{AB}(}\boldsymbol{i}\mathbf{)}$, a positive value indicates better performance for OCBIR, while a negative value indicates better performance for the shape thesaurus. The first diagram clearly indicates higher recall performance for the shape thesaurus. The second diagram indicates OCBIR achieves higher precision in the queries than the shape thesaurus.

Complete diagrams for the different query levels and query types can be found in Appendix J.

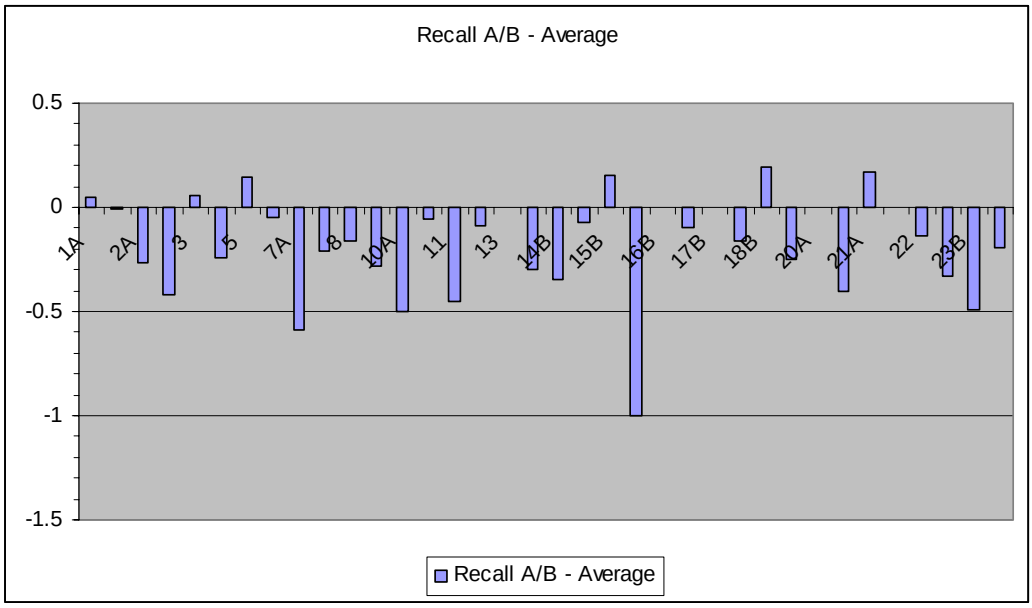


Figure 53 - $R_{AB}(i)$ histogram - All query levels, average of both query types.

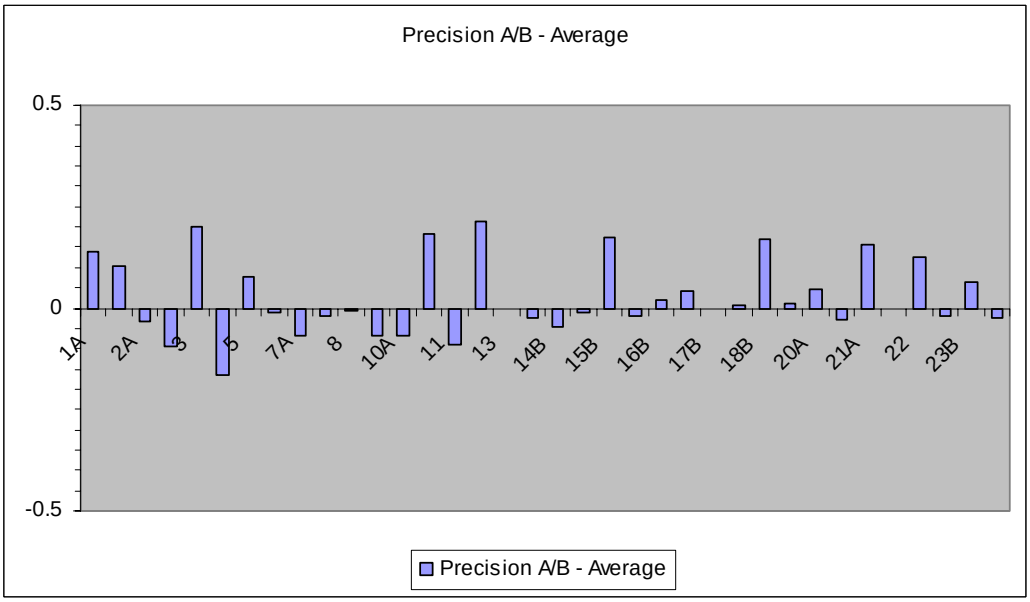


Figure 54 - $P_{AB}(i)$ histogram - All query levels, average of both query types.

Next, the single value summaries were averaged for the different query types and query levels. These averages are shown in Table 9 and Table 10, showing precision and recall respectively. From these tables, it appears that retrieval with the shape thesaurus outperforms OCBIR for all measures except precision for example-based retrieval.

Finally, the recall and precision levels achieved for each query were used as a basis for evaluating if the findings indicated by the recall / precision measurements are purely coincidental or actually significant. This is further detailed in chapter 5.3.4.

Table 9 - Average recall results for the different query levels.

Query Level	OCBIR		Shape Thesaurus	
	QBE	QBS	QBE	QBS
Generic Objects	14,13 %	11,38 %	18,38 %	20,54 %
Generic Scenes	3,67 %	3,00 %	31,78 %	26,89 %
Specific Objects and scenes	8,83 %	11,00 %	38,17 %	30,60 %
Abstract Objects, scenes and activities	23,54 %	6,69 %	29,62 %	19,23 %
Average	11,28 %	7,4 %	30,19 %	24,23 %

Table 10 - Average precision results for the different query levels.

Query Level	OCBIR		Shape Thesaurus	
	QBE	QBS	QBE	QBS
Generic Objects	13,38 %	10,50 %	6,13 %	12,38 %
Generic Scenes	10,00 %	10,13 %	6,75 %	13,25 %
Specific Objects and scenes	7,00 %	8,63 %	6,63 %	12,00 %
Abstract Objects, scenes and activities	6,50 %	7,75 %	6,38 %	10,13 %
Average	9,22 %	9,25 %	6,47 %	11,94 %

Table 11 - recall and precision for all queries.

Query Level	Query	OCBIR						Shape Thesaurus						Comparison			
		Recall			Precision			Recall			Precision			$R_{A/B}(i)$		$P_{A/B}(i)$	
		Example	Drawing	Average	Example	Drawing	Average	Example	Drawing	Average	Example	Drawing	Average	Example	Drawing	Example	Drawing
Generic Scenes	1A	0.6	0.1	0.35	0.33	0.07	0.20	0.3	0.3	0.30	0.06	0.06	0.06	0.3	-0.2	0.27	0.01
	1B	0.13	0.11	0.12	0.27	0.15	0.21	0.13	0.13	0.13	0.09	0.12	0.11	0	-0.02	0.18	0.03
	2A	0.07	0.13	0.10	0.04	0.07	0.06	0.13	0.6	0.37	0.03	0.15	0.09	-0.06	-0.47	0.01	-0.08
	2B	0	0	0.00	0	0	0.00	0.67	0.17	0.42	0.11	0.08	0.10	-0.67	-0.17	-0.11	-0.08
	3	0.19	0.11	0.15	0.36	0.13	0.25	0.04	0.15	0.10	0.02	0.07	0.05	0.15	-0.04	0.34	0.06
	4	0	0.17	0.09	0	0.3	0.15	0.13	0.52	0.33	0.16	0.47	0.32	-0.13	-0.35	-0.16	-0.17
	5	0.14	0.29	0.22	0.07	0.12	0.10	0.07	0.07	0.07	0.02	0.02	0.02	0.07	0.22	0.05	0.1
	6	0	0	0.00	0	0	0.00	0	0.09	0.05	0	0.02	0.01	0	-0.09	0	-0.02
Generic Scenes	7A	0.08	0.08	0.08	0.06	0.04	0.05	0.67	0.67	0.67	0.11	0.13	0.12	-0.59	-0.59	-0.05	-0.09
	7B	0.08	0.08	0.08	0.03	0.03	0.03	0.5	0.08	0.29	0.08	0.02	0.05	-0.42	0	-0.05	0.01
	8	0	0	0.00	0	0	0.00	0.33	0	0.17	0.01	0	0.01	-0.33	0	-0.01	0
	9	0	0.06	0.03	0	0.03	0.02	0.06	0.56	0.31	0.02	0.15	0.09	-0.06	-0.5	-0.02	-0.12
	10A	0	0	0.00	0	0	0.00	0.5	0.5	0.50	0.07	0.07	0.07	-0.5	-0.5	-0.07	-0.07
	10B	0.08	0.05	0.07	0.27	0.2	0.24	0.08	0.16	0.12	0.08	0.02	0.05	0	-0.11	0.19	0.18
	11	0	0	0.00	0	0	0.00	0.45	0.45	0.45	0.09	0.09	0.09	-0.45	-0.45	-0.09	-0.09
	12	0.09	0	0.05	0.5	0	0.25	0.27	0	0.14	0.07	0	0.04	-0.18	0	0.43	0
	13	0	0	0.00	0	0	0.00	0	0	0.00	0	0	0.00	0	0	0	0
Specific	14A	0	0	0.00	0	0	0.00	0.6	0	0.30	0.05	0	0.03	-0.6	0	-0.05	0
	14B	0	0.2	0.10	0	0.04	0.02	0.5	0.4	0.45	0.08	0.05	0.07	-0.5	-0.2	-0.08	-0.01
	15A	0	0	0.00	0	0	0.00	0.14	0	0.07	0.02	0	0.01	-0.14	0	-0.02	0
	15B	0.4	0.1	0.25	0.27	0.11	0.19	0.2	0	0.10	0.03	0	0.02	0.2	0.1	0.24	0.11
	16A	0	0	0.00	0	0	0.00	1	1	1.00	0.02	0.02	0.02	-1	-1	-0.02	-0.02
	16B	0.13	0.25	0.19	0.04	0.06	0.05	0.25	0.13	0.19	0.04	0.02	0.03	-0.12	0.12	0	0.04
Abstract	17A	0.27	0.09	0.18	0.06	0.03	0.05	0	0.55	0.28	0	0.01	0.01	0.27	-0.46	0.06	0.02
	17B	0	0	0.00	0	0	0.00	0	0	0.00	0	0	0.00	0	0	0	0
	18A	0.17	0.08	0.13	0.07	0.04	0.06	0.5	0.08	0.29	0.08	0.02	0.05	-0.33	0	-0.01	0.02
	18B	0.33	0.17	0.25	0.13	0.25	0.19	0.06	0.06	0.06	0.02	0.02	0.02	0.27	0.11	0.11	0.23
	19	0	0.2	0.10	0	0.13	0.07	0.3	0.4	0.35	0.05	0.06	0.06	-0.3	-0.2	-0.05	0.07
	20A	0.33	0.33	0.33	0.08	0.04	0.06	0.33	0.33	0.33	0.02	0.01	0.02	0	0	0.06	0.03
	20B	0	0	0.00	0	0	0.00	0.8	0	0.40	0.06	0	0.03	-0.8	0	-0.06	0
	21A	0.67	0	0.34	0.33	0	0.17	0.33	0	0.17	0.02	0	0.01	0.34	0	0.31	0
	21B	0	0	0.00	0	0	0.00	0	0	0.00	0	0	0.00	0	0	0	0
	22	0.29	0	0.15	0.33	0	0.17	0.57	0	0.29	0.08	0	0.04	-0.28	0	0.25	0
	23A	0	0	0.00	0	0	0.00	0.33	0.33	0.33	0.02	0.02	0.02	-0.33	-0.33	-0.02	-0.02
	23B	0.01	0	0.01	0.17	0	0.09	0.5	0.5	0.50	0.02	0.02	0.02	-0.49	-0.5	0.15	-0.02
	24	0	0	0.00	0	0	0.00	0.13	0.25	0.19	0.02	0.03	0.03	-0.13	-0.25	-0.02	-0.03

5.3.4 Significance Testing

The different data views presented in the previous chapter would appear to support the conclusion that the shape thesaurus outperforms OCBIR in most cases. However, in order to establish if the indicated differences are real or merely coincidence, a significance test would be needed. Several statistical tests are available for testing for significant differences between two data sets. However, as noted by van Rijsbergen (1999), there are some problematic issues with using such tests for evaluation

information retrieval systems. Statistical tests are based on assumptions about the nature of the underlying data, such as the *t-test*:

- The samples are independently and randomly drawn from the source population(s).
- That the scale of measurement for both samples has the properties of an equal interval scale.
- That the source population(s) can be reasonably supposed to have a normal distribution.

Although it might be argued that the two first assumptions hold for information retrieval evaluations, it is difficult to establish whether the source population has a normal distribution. However, for purposes of significance evaluation in this thesis, it is assumed that the data *has* a normal distribution. Consequentially, the *students-t* test is used to evaluate significance in search results. As a result, any conclusions drawn from this test should be considered under this assumption.

In order to test for significance, the following null-hypothesis was formulated:

> *Use of a shape thesaurus does not lead to any significant improvement in recall or precision.*

The null hypothesis will be discarded if can be established with a probability of 95% that the observed results are not coincidence, i.e. a 0.05 level of significance.

The null-hypothesis above refers to the total recall and precision achieved by the retrieval systems. This was measured using the “Average” column, for both OCBIR and the shape thesaurus, in Table 11. Furthermore, null-hypotheses were similarity formulated and tested for all the criterions shown in Table 7, page 73, using the data subsets defined by the different query levels in Table 11.

Table 12 shows the results from significance testing for the average (both example and image) recall results for all query levels, while Table 13 shows similar results for precision. A basic knowledge of significance testing with *student-t* test is assumed.

Table 12 - Paired two-sample t-test for average recall, both images and examples.

	Variable 1	*Variable 2*
Mean	0.26875	0.0925
Variance	0.04231625	0.01160214
Observations	36	36
Pearson Correlation	-0.169032629	
Hypothesized Mean Difference	0	
Df	35	
t Stat	4.26741	
P(T<=t) one-tail	0.00007	
t Critical one-tail	1.68957	
P(T<=t) two-tail	0.00014	
t Critical two-tail	2.03011	

Table 13 - Paired two sample t-test for average precision both images and examples.

	Variable 1	*Variable 2*
Mean	0.072916667	0.04722222
Variance	0.007379107	0.00316492
Observations	36	36
Pearson Correlation	0.185297996	
Hypothesized Mean Difference	0	
df	35	
t Stat	1.64782	
P(T<=t) one-tail	0.05417	
t Critical one-tail	1.68957	
P(T<=t) two-tail	0.10834	
t Critical two-tail	2.03011	

t Stat is the t-value of the t-test. *t Critical one-tail* is the t- threshold for a significance level of 0.05 in a one-tailed test, while *t Critical two-tail* is the similar threshold for a two-tailed test. Since the hypothesis indicates that the difference should be from retrieval with the shape thesaurus to OCBIR, we have a directional test, or a one-tailed test, and the first threshold value is used.

For the recall test results shown in Table 12, we see that the resulting t-value is indeed higher than the critical threshold value, indicating that the differences in average recall is indeed significant at the 0.05 level. In this table, the results are even significant on a 0.001 level, indicating that the findings are highly significant. However, for the precision results shown in Table 13, we see that the actual t-value is below the 0.05 threshold, indicating that the results are not significant for this level.

The t-test results for all sub-hypothesis are found in Appendix K. Table 14 shows an overview of the significance tests for the different query levels and query types. **Bold** text indicates that the results were found to be significant at the 0.05 level. “ST” and “OCBIR” indicates whether the differences were in favour of retrieval with the shape thesaurus or OCBIR.

Table 14 - Overview of T-test significance results.

	Recall		Precision	
Query Level	Example	Drawing	Example	Drawing
All Levels	**ST**	**ST**	**OCBIR**	OCBIR
Generic Objects	ST	ST	OCBIR	ST
Generic Scenes	**ST**	**ST**	OCBIR	ST
Specific	**ST**	ST	OCBIR	OCBIR
Abstract	ST	**ST**	**OCBIR**	OCBIR
Overall	**ST**		ST	

For recall, all tests indicated a difference in favour of the shape thesaurus. We also see that the tests reveal that there is a *significant* difference for both query types (QBE and QBS), as well as the average[17] recall results achieved with retrieval with the shape thesaurus. For precision, all results except QBS for generic objects and scenes were in

[17] The ”Average” recall values are based on the “Average” columns in Table 11

favour of OCBIR. Furthermore, we see that these results were found to be *significant* for example images.

These results are discussed further in chapter 6.1

5.3.5 Processing the Questionnaire

The questionnaire was not a central part of this project, and hence analysis of this was not given high priority. Furthermore, the amount of collected data is very small, and would not be very meaningful to perform a statistical analysis. The data material obtained from the questionnaire was subject to an informal analysis, attempting to identify potential interesting indications.

The first set of questions, as well as question 5 and 6, concerned how the respondents felt about finding, and drawing, query images. The answers were re-grouped according to query type and query level and averaged. The scale of the answers varies from 1 (easy) to 6 (difficult). The results are presented in Table 15, below.

The second set of questions concerned the perceived match between the query images and the images in the image collection. The answers were re-grouped similarly to the first set of questions, using the same scale. The resulting data is presented in Table 16.

In the last set of questions, the respondents could express difficulties with the different queries using their own words. Their responses were translated into English, and grouped according to the query level. Their answers were in line with their responses to the first questions.

A complete overview of the questionnaire answers can be found in Appendix I.

Table 15 - Complexity of expressing queries visually.

Queries based on generic objects	
How easy was it to find images?	1,13
How easy was it to draw images?	3,25
Queries based on generic scenes	
How easy was it to find images?	1,78
How easy was it to draw images?	3,33
Queries based on specific objects and scenes	
How easy was it to find images?	2,17
How easy was it to draw images?	3,50
Queries based on abstract objects, scenes and activities	
How easy was it to find images?	2,54
How easy was it to draw images?	3,46
All Queries (Question 5 and 6)	
How easy was it to find images?	2,33
How easy was it to draw images?	4,17

Table 16 - Perceived match between query image and collection.

Queries based on generic objects	
Match between best seed and collection	3,38
Match between best drawing and collection	3,25
Queries based on generic scenes	
Match between best seed and collection	4,67
Match between best drawing and collection	3,56
Queries based on specific objects and scenes	
Match between best seed and collection	2,33
Match between best drawing and collection	3,17
Queries based on abstract objects, scenes and activities	
Match between best seed and collection	3,23
Match between best drawing and collection	2,77

From the results from the first four questions in Table 15 , we see that the respondents generally found it easy to find good seed images. This is illustrated by the following two comments from the respondents:

> "The internet is swamped by dolphins; there are a lot to choose from." *(Question 7, Respondent 6, Query 1)*

> "Searched on bird/sea/sky, then on 'flying seagull', and found a few relevant images. Easy to find bird, but difficult to find a flying bird with a different background". *(Question 7, Respondent 3, Query 3)*

We also see that drawing images grew slightly more difficult with the increased semantic complexity of the queries. This is illustrated further by this respondents' comment:

> "A little more difficult, because of the "surface" restriction. "How should they be on the surface? Some jumps, some, like the blue whale, just showed their back. Difficult to choose an image." *(Question 7, Respondent 3, Query 2)*

Furthermore, the respondents found it generally more difficult to draw their own images than using example images. This is reflected in both the average scores for questions 1 and 2, as well as their average opinions in questions 5 and 6. An answer from respondent 5 gives a good description of some of the difficulties using drawings to express the queries:

> "Generally when drawing, my difficulties were:
> - Paint shop Pro was for me an unfamiliar tool for processing pictures.
> - Using the PC mouse in drawing in stead of pencil was very difficult, possibly due to the fact that the movement of the hand differ very much from holding a pencil. -> less control of the result.
> - It is difficult to establish (for a person not trained in drawing) what the main features of an object is. I did for instance forget that the whale also have a very distinguishable fin on its back when drawing the whale" *(Question 8, respondent 5, Query 1)*

Although the data is very limited, the results seem to indicate that users prefer Query-By-Example over Query-By-Sketch, and that they find it relatively easy to express some types of queries visually. However, this needs to be evaluated further in a more thorough study.

The questions regarding the perceived match between the seed images and the image collection (questions 3 and 4) were not given any particular evaluation. They were included in order to present the opportunity to discover if there were any discrepancies between what the users *perceived* to be the best match between their seed images and the image collection. However, as this was not a very central part of the research project, very little effort was spent towards evaluating this.

Nevertheless, one potentially interesting observation is worth mentioning. While questions 1, 2, 5 and 6 indicated that the users generally found it easier to *find* images than to create images themselves, the results from questions 3 and 4 seems to indicate that the best matches were achieved through the drawn images. However, the data material is small and the results have not been given a thorough analysis, and it is not possible to present any conclusive findings from these results.

Finally, several variables were not given consideration in this questionnaire, such as any differences in results based on sex, age, computer skills or experience with image retrieval. It is believed that such an evaluation might reveal some interesting findings.

The results from the questionnaire are not given any further comments in this thesis. As a concluding note, the brief evaluation of the results indicates that it might be very interesting follow up these questions in further studies.

6 Evaluation of Results and Conclusion

The focus of this research project, as stated in the research question, has been to examine if a Shape Thesaurus can be used to significantly improve the search capabilities of an image retrieval system. The goal was to implement a working prototype of a shape thesaurus, and compare image retrieval with a shape thesaurus to an image retrieval system based solely on low-level feature comparison techniques.

With the implementation of the shape thesaurus in the VORTEX prototype, we have seen that implementing a Shape Thesaurus is feasible, even if the implementation created in this project has several limitations. We have also seen that the prototype has been tested and compared to a pure syntactic feature based image retrieval system. However, the final question remains – is it possible to give an answer to the research question and hypothesis based on the implemented prototype and the collected data?

6.1 Hypothesis Evaluation

6.1.1 Verification / Falsification of Hypothesis

We recall that the main hypothesis tested in this project was

> *A system that utilizes a thesaurus for shapes, will lead to a significant improvement in recall / precision results over a system based on syntactical feature comparison.*

The hypothesis consists of three main components. A *Shape Thesaurus*, *search results in recall / precision* and a *significant improvement*. The *Shape Thesaurus* has been implemented as a part of the VORTEX prototype, and subsequently acts as a representative of a shape thesaurus. The *search results in recall / precision* were described in chapter 5.3.3 and can be found in Appendices H and J.

First of all, we recall that the recall / precision curves indicated that image retrieval with the shape thesaurus was slightly better for QBE, and seemed very much better with QBS (Figure 52, page 85). These findings were partly supported by the recall and precision histograms, indicating that *recall* was found to be considerably better for retrieval with the shape thesaurus, while precision was somewhat more uncertain (Figure 53 and Figure 54, page 87). Together, these measurements indicate two things:

- Recall was higher with the shape thesaurus than OCBIR in all cases
- Precision was almost equal between the two systems, slightly in favour of OCBIR.

Finally, the *significance of the results* was measured in chapter 5.3.4. The results of the significance testing indicated that:

- *Average recall* was significantly better with the shape thesaurus than OCBIR
- *Average precision* was *not* significantly better with the shape thesaurus than with OCBIR

- *Average precision* was *not* significantly better with OCBIR than with the shape thesaurus

Based on this, it is possible to propose the following answer to the central hypothesis:

> *A Shape Thesaurus, as implemented in the VORTEX prototype, is able to achieve significantly better recall results over a system based on syntactical feature comparison, as represented by OCBIR functionality*.

However, this answer is based on several assumptions, and does not address all the issues raised during the experiment. We need to evaluate these assumptions and issues before it is possible to present an answer to the central research question posed in this project.

6.1.2 Hypothesis and Search Results Breakdown

The conclusions drawn in the previous chapter were based on the overall and average retrieval results achieved by retrieval with a shape thesaurus and OCBIR. However, one of the goals in the project was to evaluate the performance of the two algorithms were for the different levels of semantic content, described in Table 7 (page 73).

Measurement of recall and precision for all these levels were made with each of the different measurement tools described in chapters 5.3.3 and 5.3.4. However, during the analysis it became evident that it was difficult to say anything certain about the different levels. The primary reason for this lies in the basic mechanisms of the shape thesaurus. These are focused at identifying *objects* and not scenes, activities or meanings. If the retrieval algorithm manages to identify an object, images containing similar objects are retrieved, not images with similar meanings, activities or even objects engaged in certain activities. As a result, I decided not to make any conclusions concerning the two algorithms' performance on the different levels of semantic complexity.

However, most of the queries used in the project were centred around visual objects (Dolphins, birds, divers and so on), and even if executing the queries did not let us evaluate the different semantic levels, they presented us with a large data material for evaluating the main focus of the shape thesaurus; to identify different variations in visual objects, and retrieve different depictions of these objects. It is believed that his is reflected in the overall performance of the two algorithms, as presented in the previous chapter.

Now, let us consider the differences between QBE and QBS. The recall and precision curves seems to indicate that the largest difference between image retrieval with the shape thesaurus and OCBIR are in the case of QBS. The recall / precision curve lies considerably higher here, while they are more similar in the case of QBE. This is indicated further by the significance testing (Table 14, page 91).We see that the differences were found to be significant for QBS and not significant at the 0.05 level for QBE. This seems to indicate that image retrieval assisted by a shape thesaurus is better suited for QBS than QBE.

One likely reason for this result is the fact that the thesaurus shape templates and the drawings created for the QBS searches were approximately of the same size (400 by

400 pixels) and format (primarily black line drawings on a white background). This is quite interesting, as it indicates that the shape thesaurus might be best used in situations where users *draw* their queries, rather than using existing images as a basis for query-by-example.

6.2 Evaluation of the Approach

6.2.1 Reliability and Validity

Ringdal (2001) describes the following model for measuring reliability and validity:

V = true value + measuring error

where *V* is a measured value. The value of *V* is determined by two components; true value and measuring errors. There are two types of measuring errors; *random measuring errors* and *systematic measuring errors*. The former concerns data reliability, while the latter concerns the validity of the data.

The *Reliability* of an experiment describes if repeated assessment with the same measuring tool provides similar results. This is decided by *how* the measurements have been made. The *Validity* of an experiment describes if we measure what we actually want to measure; if the data collected is relevant to the research question. Validity concerns *what* we measure. High validity is dependent of high reliability. Validity for experiments can further be detailed into *inner* and *outer* validity. The former describes if it is possible to determine that a given treatment X causes the response Y, while the latter describes if it is possible to generalize the results of the experiment.

The measuring tool used in this experiment is measurement of Recall and Precision. The measurements were based on comparing the relevant images as determined by the respondents, to results obtained by the two different retrieval systems based on the search images submitted by the respondents. Reliability in this experiment can be jeopardized by errors occurring while executing the queries and calculating the recall and precision values. Examples of such can be:

- Using wrong seed image as query input
- Errors when copying query results from the interface into worksheets for calculation
- Errors when calculating average recall and precision scores

One possible error source is that the shape thesaurus retrieval algorithm sometimes retrieved the same image twice, and the duplicates were not removed in the final list. However, these were not taken into consideration when finding relevant images.

As it is human to error, it is very difficult to avoid such errors completely. However, strict precautions were taken to avoid such errors, and the data and the different calculations and were checked several times in order to ensure their correctness.

As a result, it is assumed that there is a sufficiently high degree of reliability in the data material. However, as noted in chapter 5.3.4, this is a conjecture based on the

assumption that the underlying data material has a *normal distribution*. Any conclusions based on the significance testing must be considered under this reservation.

Next, several different factors influence the validity of the experiment. The first and most obvious question is – how does the VORTEX prototype reflect the Shape Thesaurus framework proposed in this thesis? If the prototype cannot be said to represent the described framework, it is at best problematic to claim that the experimental results can be used to answer the research question and hypothesis. Additionally, what can be said concerning the use of OCBIR as a measure of a standard feature-based image retrieval system? These are issues related to the *outer validity* of the experiment.

Furthermore, can we be sure that the observed results are an effect of using the planned functionality of the shape thesaurus, or are there other causes? This is an issue related to the *inner validity* if the experiment.

Additionally, are there any other reservations that should be made regarding the structure and different components of the prototype? While this might not be directly related to the research question and hypothesis, it is relevant for the outer validity of the whole research project. If there are difficulties or problems with the critical functions in the thesaurus, are these specific to VORTEX or might there be similar problems in other implementations? This goes to the *outer validity* of the proposed framework.

Finally, we might question if recall and precision are a suitable tool for evaluation of the proposed framework. This is not directly related to the research question and hypothesis used in this project, as they are actually based on using this measurement tool for comparison. However, one of the main motivations stated for described framework has been to improve semantic image retrieval. Consequently, this is relevant to the *outer validity* of the research project as a whole.

6.2.2 VORTEX and the Shape Thesaurus

How does the shape thesaurus implemented in the VORTEX prototype reflect the Shape Thesaurus described in chapter 3, and what limitations and problematic issues are present in the prototype? The shape thesaurus in the VORTEX prototype was created as one possible implementation of the proposed framework. Unless it can be shown that the implemented prototype sufficiently reflects the shape thesaurus, the experimental results gained from testing the prototype cannot be used to determine the usefulness of the proposed framework.

As stated in chapter 4.1.1, only the core functionality required by the *Shape Thesaurus* framework was included in the VORTEX Prototype. First of all we have the four components identified by the *Shape Thesaurus* definition:

> *(1) A precompiled list of important shapes representing visual objects in a given domain of knowledge*
> *(2) feature descriptors describing these shapes*
> *(3) a textual / semantic description of these shapes*
> *(4) for each shape, a set of related shapes.*

Furthermore we have the *shape recognition mechanism*. Combined, these five elements represent the core functionality of the shape thesaurus.

If we compare this to the shape thesaurus implemented in the VORTEX prototype, we see that all these are included:

1. The list of *visual objects and shapes*, as well as the *semantic descriptions* of these are included in the *Thesaurus_Object* table
2. Feature descriptors are implemented through the signatures representing the shape templates, included in the OrdImageSignature variable *shape_sig* in the *Thesaurus_representation* table.
3. The different thesaurus relationships are implemented in either through the *is_related* table, or with different shape descriptors for the *variant-of* relationships.
4. The *IMGSimilar* and *IMGScore* functions of the *OrdSys* class contain functionality for shape recognition.

Finally, the shape thesaurus functionality in the VORTEX prototype includes basic support for query expansion and ranking the images in the result set.

Based on this, I propose to accept the premise that the implemented shape thesaurus might act as a *good enough* representation of a Shape Thesaurus to act as an indication of the capabilities of the framework.

Support for other query types (Text to Visual and Visual to Text) should be relatively easy to implement, as they are based on using the same principles as the visual to visual query implemented. Furthermore, support for user specification of query expansion and result ranking is also believed to be reasonably easy to implement in a user interface. However, this should be tested before any conclusive statements are put forward.

Next, there are some issues with the prototype that should be discussed. First of all, the shape descriptors and the shape recognition mechanism used are likely to be far from optimal. First of all, as noted in chapter 2.5.4, the rigid structure of basic shape templates has several limitations, and *deformable* shape prototypes might have been a better choice. Unfortunately, implementing a system based on deformable prototypes would take too much time and effort to be viable within the time frame available for this project.

Furthermore, the inner workings of both the feature descriptors (OrdImageSignature) and the actual similarity functions (IMGSimilar and IMGScore) used in the prototype are hidden, which makes the use of these problematic. First of all, I would not be able to have any control over the segmentation process. In fact, very little information is available on how the OrdImageSignature performs image segmentation. It is therefore difficult to know if OrdImageSignature is actually capable of segmenting out distinct objects in an image. According to (Ward 2001), OrdImageSignature is capable of segmenting an image into different regions based on colours, and each of these regions are analyzed with respect to colour, texture and shape properties. Nevertheless, it was difficult to establish the capabilities of this segmentation through

literature, and there would be no time to evaluate this within the timeframe of this project.

Furthermore, I would have no control over the implementation and usage of the similarity functions used by O9i.The OrdImageSignature class contains a set of different algorithms and similarity functions, but these are hidden. It is likely that the algorithms used are suited for a very general application area, and that some sort of customization would be appropriate before they could be fully utilized in the context of a shape thesaurus. Moreover, it has proved difficult to establish the capabilities of the built in similarity functions. (Hove 2003) presented some of the similarity functions, which seemed to indicate that it worked satisfactorily at least for the *colour* property. However, no conclusive testing was performed, and there would not be any time to evaluate this properly within the time frame of this project.

As a result of this, it is difficult to say anything conclusive about the quality of the descriptors and similarity functions used in this project. Some indication is reflected by the low overall recall and precision achieved by the two systems. However, as the same similarity functions and descriptors were used in both retrieval algorithms. The existence of the shape thesaurus is the main differentiation between the two algorithms. This does to some degree negate the problems with the shape descriptors and the similarity functions for the purposes of this project.

Some time was spent searching for alternate feature descriptors. Different approaches were evaluated, such as the *image toolbox* in *MathLab* and classes from the *Open Computer Vision* society. Furthermore, some effort was made towards using these descriptors in conjunction with an Artificial Neural Network, created in *BrainMaker*. However, due to the limited time available and several issues with combining the neural network to the feature descriptors, this approach had to be abandoned.

Finally, the functionality of the thesaurus implemented in the VORTEX prototype is based on *global image matching*; the *entire* seed image is compared to the shape templates. While this works well in cases where:

- a QBE seed image contains a depiction of *one* object, or
- a QBS seed image containing a single drawing of one visual object.

Even though the *OrdImageSignature* and the *IMGSimilar* function is supposedly capable of object segmentation, the thesaurus objects are described through single images containing shapes describing the thesaurus objects. Because of this, object identification will only work if the query images contain single, clearly defined objects.

Despite these limitations, it is believed that the results obtained in the experiment are interesting, as the same descriptors and similarity functions are used in both OCBIR and image retrieval with the shape thesaurus. It has been shown that the shape thesaurus in the VORTEX prototype can improve image retrieval, even with these limitations. However, it has been shown that the Shape Thesaurus framework, through the implementation in the VORTEX prototype, is capable of achieving better search results even under these limitations. This indicates that even better results might be

achieved using known and customizable shape descriptors and shape similarity functions, and should be tested further.

Furthermore, the links between thesaurus objects and the image collection has been made manually, which puts some further limitations on the prototype, particularly in respects to the problems of *volume and subjectivity*. Although it would be possible to use the similarity functions to identify thesaurus objects in the image collection, I decided against this for the purposes of the prototype. Because of the limitations in the similarity functions, I found it likely these would result in a high number of incorrect links.

Table 17 below shows the distribution between linked and non-linked of retrieved relevant images retrieved using the Shape Thesaurus. The queries shown are the queries where the shape thesaurus retrieval algorithm managed near-identification of the objects in the query image.

Table 17 - Overview of image retrieval - linked vs. non linked images.

	Retrieval statistics		Distribution of linked vs. non linked retrieved images	
Query	Total Relevant	Relevant retrieved	Linked	Non Linked
2A Drawn	15	8	7	1
2B Drawn	12	8	7	1
7A Example	12	8	6	2
7A Drawn	12	7	5	2
7B Example	12	6	5	1
10A Example	10	5	4	1
10A Drawn	10	5	4	1
11 Example	11	5	4	1
11 Drawn	11	5	4	1
14A Example	5	3	3	0
14B Example	10	5	4	1
14B Drawn	10	4	1	3
17A Drawn	11	6	6	0
18A Example	12	6	4	2
19 Example	10	3	2	1
19 Drawn	10	4	3	1
20B Example	5	1	1	0
22 Example	7	2	1	1

From the table, we see that most of the relevant images retrieved were based on the links between the thesaurus object and the image collection. Consequently, we do not have any evidence to support the claim that the Shape Thesaurus might alleviate the problems of *Volume and Subjectivity*. The results must be observed under the reservation that mapping between the thesaurus and image collection has been created manually.

Finally, we need to question the use of *shape templates* as a basis for object description. The shape templates used in this project took about 2 week's full time work to assemble and create. It is believed that a working system ought to contain a

larger set of templates, as those included here clearly limits the recognizable shapes. Furthermore, in order for the shape thesaurus to be useful, it is likely that a larger thesaurus would be needed. This indicates that creating a full sized thesaurus for a particular domain might prove to be a major undertaking. However, once a working thesaurus has been implemented, it ought to be possible to adapt it different image retrieval systems in the same domain, assuming the image descriptors are similar or similar image descriptors can be made from the images. Consequently, creating a thesaurus is likely to be a one-time job, as opposed to creating a full set of textual descriptors for an image collection.

6.2.3 Use of Recall and Precision as a Measurement Tool

One final concern with this experiment is if recall and precision is an optimal tool for evaluating a tool such as the shape thesaurus. While this measurement tool presents us with a measure for the quality of the result sets, and indicates that the inclusion of the shape thesaurus has a positive effect on the quality, it does not actually say anything about how good the shape thesaurus mechanisms are at actual object recognition.

As a result, the data material collected in this experiment does not present us with any clear evidence to support any claim as to how *good* the implemented shape thesaurus is at recognizing objects. Further experiments using other measurement tools should be performed in order to gain a clearer understanding of this issue.

6.3 Conclusions from the Research Project

Finally, what final conclusions can be drawn from this research project? Based on the discussions in the previous chapter, it is believed that we have a relatively high degree of *reliability*. There seems to be few sources that might jeopardize the reliability, as a high degree of control has been held over the measurement tools.

Furthermore, we may assume that the *internal* validity of the experiment with respect to the research question and hypothesis is relatively high. The only difference between the two algorithms is the inclusion of a shape thesaurus prototype. The respondents, query set and set measuring tools are identical for the two retrieval algorithms, and the attribute of interest are the final recall and precision measurements. Accordingly, the observed differences in recall and precision between the two systems are most likely to occur from the inclusion of the shape thesaurus.

The main uncertainty regarding the internal validity is the concern about recall and precision as a measurement tool. As we do not have any clear evidence as to the object identification capabilities of the shape thesaurus, it is difficult to say anything conclusive towards the shape thesaurus' capabilities of identifying the semantic content in an image. However, while this gives us less data about these capabilities, it is believed that this does not actually reduce the internal validity of the experiment, as the data collected and observed *are relevant to the research question*. However, this issue raises the concern whether the experimental design and measurement tools used in this project were well suited for an evaluation of the shape thesaurus. It is felt that future studies of the shape thesaurus, and similar image approaches to image retrieval, should be based on other, potentially more suited measurement tools.

Concerning the *external validity* of the experiment, this might not be considered as high as the reliability and internal validity. We have seen that there are several issues regarding the underlying mechanisms, such as little control over the shape descriptors and the similarity functions, and the fact that retrieval is mainly based on pre-defined links between the thesaurus and the image collection. This might make the results of the experiment difficult to generalize image retrieval in general, and the proposed Shape Thesaurus framework in particular.

We then return to the central research question posed in this thesis;

> *Can recall / precision measures for an image retrieval system be significantly improved by utilizing a thesaurus for shapes?*

The experiment has shown that it is, under several limitations, possible to improve image recall by using a thesaurus for shapes. However, it is not felt that the results from this project are strong enough to give a conclusive answer to this question. Furthermore, due to the experimental design and the measurement tools chosen, it is difficult to give a justified answer to how good a shape thesaurus is at identifying visual objects.

Nevertheless, it is felt that the experimental results have shown that the proposed framework is worthy of further development and evaluation. Accordingly, the project described in this thesis might be regarded as a pilot study, and that the results here might prove useful for further explorations into the possibilities of the proposed framework.

6.4 Future Research

6.4.1 Improving the VORTEX Implementation

The discussions in the previous chapter revealed some limitations in the shape thesaurus in VORTEX prototype that reduces its usefulness as a representative of the framework presented in chapter 3. In order to gain a higher understanding of the potential of a Shape Thesaurus, several components of the current VORTEX implementation ought to be expanded.

First of all, OrdImageSignature and its related similarity functions should be replaced. The use of these puts several restrictions on the prototype and the experimental results. It is believed that a solution based on *deformable shape prototypes* present a better solution than the rigid templates used in the current implementation.

Furthermore, as mentioned in the previous chapter, some experimentation was done with using an Artificial Neural Network (ANN) as a tool for identifying and recognizing shapes present in an image. ANNs have been successfully used in other image retrieval applications, and it is believed that this might prove useful combined with the mechanisms of a shape thesaurus, possibly in combination with deformable shape templates.

Next, we saw that the shape thesaurus retrieval algorithm in the VORTEX prototype primarily returned images which were pre-linked to the thesaurus objects. This connection should be further explored. First of all, retrieval from non-linked images ought to be increased. This is related to the similarity functions and descriptors discussed in the previous paragraph. In Table 3, page 13, we identified several possible approaches to mapping the thesaurus to an image collection. It is believed that a combination of these approaches could prove to be a good approach when implementing a shape thesaurus framework. On one hand, it is unlikely that all images will be completely analyzed and all content successfully identified, unless the collection is small. This again suggests that a similarity search is necessary. However, having established relationships between some images and the terms will give a faster, and perhaps better, search result than a similarity search alone. Further research and experiments should be done to examine and evaluate the different methods described here, as well as discover other approaches to this. First, better shape identification will definitely help towards this goal. Furthermore, we should examine if user interaction could be used, for example in a relevance-feedback structure. One possible approach to this is currently being evaluated in another thesis project “Continuously Updated Metadata” (Translated title). The results from this thesis might prove to be interesting for thesaurus mapping.

Independently of how the terms are linked to image content, it is likely that there will be some erroneous links. Incorrect links will result in inferior search results. Accordingly, some sort of error-correction mechanism could be provided. One possible approach would be to use a relevance / feedback cycle to discover and correct erroneous links. If users could be encouraged to mark up relevant and irrelevant images, this information could be used to find and correct incorrect links, as well as establish new links between images and thesaurus terms. Further research should be done to explore this avenue as well as examine other possibilities.

The ability for a shape thesaurus to assist in visible object identification has not been evaluated in this thesis. This ought to be examined further, as it will further increase our understanding of the usefulness of the Shape Thesaurus.

Finally, the Shape Thesaurus functionality currently missing in the VORTEX prototype should be implemented. As this functionality has not been tested, it is impossible to determine their usefulness. A complete testing of a better Shape Thesaurus will also increase our understanding of its capabilities.

6.4.2 60

6.4.3 Further Research in Image Retrieval

The answers in the questionnaire revealed one interesting trend. The respondents found it easier to express queries visually by using example images rather than drawing the images themselves. In addition, they expressed some difficulties in expressing actions and activities visually. One interesting future research project would be to study how we can assist users in expressing complex queries towards visual data. Can users be helped either through visualizations of their textual queries, or through interface tools developed to assist them in drawing image? Furthermore, is it possible to utilize the Shape Thesaurus framework, or similar existing solutions towards this goal?

The framework described in this thesis is primarily aimed at *one collection at a time*. However, one of the major challenges facing image retrieval in particular, as well as information retrieval in general, is retrieval from multiple sources. It would be interesting to evaluate if the framework proposed in this thesis, or similar approaches, could be adapted to retrieval from multiple sources. One of the largest hindrances to this might be that images are likely to have differing feature descriptors, and in many cases no descriptors *at all*. Further studies should be done in order to evaluate how these challenges can be overcome.

Finally, while the results from the questionnaire were only given a cursory evaluation, the observed results indicate that a further research should be done in this direction.

7 Bibliography and References

Baeza-Yates, R. and B. Ribeiro-Neto (1999). Modern Information Retrieval. New York, Addison-Wesley.

Beichner, R. J. and R. A. Serway (2002). Physics For Scientists and Engineers, Thomson Leaning, Inc.

Bouet, M., A. Khenchaf, et al. (1999). Shape Representation for Image Retrieval. ACM MM'99, Orlando, Florida, ACM.

Carlin, M. (2000). Improving the Performance of Shape Similarity Retrieval Systems. Department of Informatics, The Faculty of Mathematics and Natural Sciences,. Oslo, University of Oslo.

Carson, C. (1997). Region Based Image Query. Proceedings of IEEEE, CVPR '97, Workshop on Content-Based Access of Image and Video Libraries, Santa Barbara, Calefornia.

Colombo, C. and A. Del Bimbo (2002). Visible Image Retrieval. Search and Retrieval of Digital Imagery. V. Castelli and L. D. Bergman, John Wiley & Sons, Inc**:** 11-33.

Dobie, M., R. Tansley, et al. (1998). A Flexible Architecture for Content and Concept Based Multimedia Information Exploration. Challenge of Image Retrieval, Newcastle.

Dryden, I. L. and K. V. Mardia (1998). Statistical Shape Analysis. Chichester, John Wiley & Sons.

Dunckley, L. (2003). Multimedia Databases. London, Addison-Wesley.

Eakins, J. P. (1996). Automatic Image Retrieval - Are we getting anywhere. Third International Conference on Electronic Library and Visual Information Research (Elvira3), De Montfort University, Milton Keynes.

Eakins, J. P. and M. E. Graham (1999). Content Based Image Retrieval: A report to the JISC Technology Applications Program. Newcastle, Inst. for Image Data Research, Univ. of Northumbria.

Engan, K., K. F. Fretheim, et al. (2004). Compression of Digital Mammographs Without Interering With A System for Automatic Detection of Microcalcifications - Mammoscan µCaD. Norwegian Conference on Image Processing and Pattern Recognition, Stavanger, NOBIM.

Flickner, Sawhney, et al. (1995). "Query by Image and Video Content: The QBIC System." IEEE Computer **28**(9): 23-32.

Foskett, D. J. (1997). Thesaurus. Readings in information retrieval. K. S. Jones and P. Willet, Morgan Kaufmann Publishers**:** 111-134.

Gould, I. H. (1971). IFIP Guide to Concepts and Terms in Data Processing. London, North-Holland Publishing Company.

Guros, L. (2004) Retrieving data from the "ORDImageSignature Field". Accessed: January 6th, 2004. Available Online: http://forums.oracle.com/forums/thread.jsp?forum=78&thread=222477&message=617161&q=7369676e6174757265#617161

Hillman, D. (2003) Using Dublin Core. Accessed: March 2nd, 2004. Available Online: http://dublincore.org/documents/usageguide/

Hove, L.-J. (2003). Characteristics of Oracle interMedia OrdImage. J. Nordbotten. Bergen, Institute for Information- and Media Sciences.

Hove, L.-J. (2004). Improving Content Based Image Retrieval with a Thesaurus for Shapes. Norwegian Conference on Image Processing and Pattern Recognition, Stavanger, NOBIM.

Huang, T. S. and Y. Rui (1999). "Image Retrieval: Current Techniques, Promising Directions And Open Issues." Journal of Visual Communication and Image Representation **10**(4): 39-62.

Jacobson, I., G. Booch, et al. (1998). The Unified Software Development Process, Addison Wesley Longmand, Inc.

Jaimes, A. and S.-F. Chang (2002). Concepts and Techniques for Indexing Visual Semantics. Image Databases: Search and Retrieval of Digital Imagery. V. Castelli and L. D. Bergman, John Wiley & Sons, Inc.**:** 497-565.

Kimia, B. B. (2002). Shape Representation for Image Retrieval. Image Databases: Search and retrieval of Digital Imagery. V. Castelli and L. D. Bergman, John Wiley & Sons, Inc**:** 345-372.

Li, Y. and C.-C. J. Kuo (2002). Introduction to Content-Based Image Retrieval - Overview of key techniques. Image Databases: Search and retrieval of Digital Imagery. V. Castelli and L. D. Bergman, John Wiley & Sons, Inc**:** 261-284.

Lim, J.-H. (1999). Learnable Visual Keywords for Image Classification. 4th International Conference on Digital Libraries, Berkeley, California, United States, ACM Press.

Lim, J.-H. (2000). "Photograph Retrieval and Classification by Visual Keywords and Thesurus." New Generation Computing **18**(2): 147-156.

Lu, G. (1999). Multimedia Database Management Systems. Norwood, Artech House INC.

Manjunath, B. S. and W.-Y. Ma (2002). Texture Features for Image Retrieval. Image Databases: Search and Retrieval of Digital Imagery. V. Castelli and R. Baeza-Yates, John Wiley & Sons, Inc**:** 313-344.

Martínez, J. M. (2003) MPEG-7 Overview. Accessed: March 2nd, 2004. Available Online: http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm

Nordbotten, J. (2002) Virtual Exhibits on Demand. Accessed: September 13th, 2004. Available Online: http://nordbotten.ifi.uib.no/VirtualMuseum/VMwebSite/VEDweb-site.htm

Nordbotten, J. (2004). Advanced Data Management. Bergen.

Ringdal, K. (2001). Enhet og Mangfold: Samfunnsvitenskapelig forskning og kvantitativ metode. Bergen, Fagbokforlaget.

Rowley, H., S. Baluja, et al. (1998). "Neural Network Based Face Recognition." Pattern Analysis and Machine Intelligence(20): 23-38.

Rui, Y., T. S. Huang, et al. (1998). Relevance Feedback Techniques in Interactive Content-Based Image Retrieval. Storage and Retrieval for Image and Video Databases**:** 25-36.

Santini, S. and R. Jain (1997). "The Graphical Specification of Similarity Queries." Journal of Visual Languages and Computing **7**: 403-421.

Sclaroff, S. (1997). "Deformable Prototypes for Encoding Shape Categories in Image Databases." Pattern Recognition **30**(4).

Sclaroff, S. and L. Liu (2000). Index Trees for Efficient Deformable Shape-based Retrieval. IEEE Workshop on Content-Based Acces of Image and Video Libraries, Hilton Head, SC, IEEE.

Sclaroff, S. and L. Liu (2001). "Deformable Shape Detection and Description Via Model-Based Region Grouping." IEEE Transactions on Pattern Analysis and Machine Intelligence **23**(5): 475-789.

Smith, J. R. (2002). Color for Image Retrieval. Image Databases: Search and Retrieval of Digital Imagery. V. Castelli and R. Baeza-Yates, John Wiley & Sons**:** 285-311.

Tansley, R., C. Bird, et al. (2000). Autoamting the linking of Content and Concept. ACM Multimedia 2000, Los Angeles, California, ACM.

van Rijsbergen, C. J. (1999) Information Retrieval. Accessed: September 24th, 2004. Available Online: http://www.dcs.gla.ac.uk/~iain/keith/index.htm

Ward, R. (2001) Oracle interMedia User's Guide and Reference. Accessed: August 1st, 2004. Available Online: http://download-west.oracle.com/docs/cd/B10501_01/appdev.920/a88786/title.htm

Woodrow, R. (1999) Iconography, Iconology. Accessed: June 6th, 2004. Available Online: http://www.newcastle.edu.au/discipline/fine-art/theory/analysis/panofsky.htm

Zloof, M. (1975). Query By Example. AFIPS, Anaheim, AFIPS Press.

Østbye, H. and T. Schwebs (1999). Media i Samfunnet, Det Norske Samlaget.